Uniwersytet Mikołaja Kopernika w Toruniu

Wydział Nauk Ekonomicznych i Zarządzania

Jakub Borko

nr albumu 273912

Wykorzystanie metod scoringowych do oceny zdolności kredytowej klientów Home Credit Group

Praca magisterska

kierunek: ekonomia

Opiekun pracy dyplomowej

Prof. UMK dr hab. Elżbieta Szulc

Katedra Ekonometrii i Statystyki

Toruń 2020

Spis treści

Wstęp

Od początku swego funkcjonowania banki starają się wypracować metodę pozwalającą ograniczyć ryzyko i koszty związane z udzielaniem kredytów. Zarządzanie ryzykiem stało się kluczowym obszarem działalności banków, decydującym o bezpieczeństwie i zyskowności biznesu. Jednym z głównych etapów procesu zarządzania ryzykiem jest identyfikacja i kwantyfikacja ryzyka. Kwantyfikacja przeprowadzana jest z wykorzystaniem różnych metod w zależności od rodzaju zidentyfikowanych ryzyk oraz poziomu zaangażowania banku w określone rodzaje transakcji. Banki do szacowania ryzyka stosują rozmaite metody od prostych opisowych po zaawansowane analizy portfela kredytowego. Jedną z najbardziej efektywnych metod jest scoring kredytowy, który można scharakteryzować jako ilościowe narzędzie wykorzystywane do oceny zdolności kredytowej potencjalnych kredytobiorców. Jest to statystyczno-matematyczny instrument umożliwiający bankowi przyporządkowanie klienta do odpowiedniej kategorii ryzyka.

Ilościowe podejście do zarządzania ryzykiem kredytowym diametralnie przybrało na znaczeniu w konsekwencji wybuchu kryzysu finansowego z 2007 roku, zapoczątkowanego na rynku kredytów hipotecznych w USA. Od tego momentu zaobserwować możemy szczególne zainteresowanie ekspertów, jak i naukowców rozwojem teorii i modeli ryzyka. Poprzez odpowiednie zautomatyzowanie i zmatematyzowanie tej dziedziny, zarządzanie ryzykiem w sektorze bankowym staje się coraz bardziej efektywne i innowacyjne. Obecnie dzięki zaawansowanym technologiom jesteśmy w stanie mierzyć ryzyko i tworzyć rozwiązania pomagające jeszcze bardziej efektywnie je kontrolować. Tworzone jest specjalistyczne oprogramowanie dedykowane dla instytucji finansowych, które gromadzi informacje z ogromnych hurtowni danych, by na ich podstawie analizować profil ryzyka klienta.

W polskiej literaturze ekonomicznej występuje deficyt prac, które zarówno pod względem teoretycznym, jak i praktycznym analizowałyby istotę oraz procedurę ilościowego modelowania ryzyka kredytowego. Szczególnie tematyka modeli scoringowych budowanych z wykorzystaniem nowoczesnych metod eksploracji danych rzadko jest podejmowana przez ekonomistów. Charakterystyczne staje się, że intensyfikowanie badań w obszarze ryzyka kredytowego następuje w okresach globalnych przeobrażeń gospodarczych oraz wraz z rozwojem nowych technologii, pozwalających skuteczniej modelować ryzyko kredytowe.

Celem niniejszej pracy jest przedstawienie metodologicznych aspektów zastosowania scoringu do oceny zdolności kredytowej dla klientów indywidualnych Home Credit Group w

kontekście wykorzystania metod ilościowych. Na przykładach omówiono główne etapy budowy systemu skoringowego: fazę projektowania systemu, fazę wdrażania wraz z etapem walidacji oraz w ujęciu teoretycznym fazę monitoringu jego stabilności. W tym celu, postanowiono zbudować 7 modeli scoringowych z wykorzystaniem następujących metod klasyfikacji: regresji logistycznej, drzew decyzyjnych, lasów losowych, drzew wzmacnianych gredientowo, maszyn wektorów wspierających, sieci neuronowych oraz algorytmu XGBoost. Postawiono hipotezę, że metody klasyfikacji wywodzące się z technik data mining będą bardziej skutecznie określały właściwą kategorię ryzyka kredytobiorcy niż metody statystyczne. Ponadto założono, że spośród metod eksploracji danych najlepsze własności klasyfikacji wykażą metody wykorzystujące gradientowe wzmacnianie. Dodatkowo w ramach analizy struktury najlepszego modelu postanowiono zbadać jakie cechy kredytobiorcy mają największy wpływ na ocenę zdolności kredytowej.

Praca składa się z czterech rozdziałów. W pierwszym rozdziale zostały przedstawione teoretyczne aspekty związane z ryzkiem kredytowym i sposobami zarządzania tym ryzykiem. Opisano także credit scoring zarówno od strony definicyjnej, jak i w ujęciu historycznym. Ponadto przedstawiono całą procedurę budowy systemu credit scoringowego.

W rozdziale drugim skupiono się na opisie metod klasyfikacji wykorzystywanych w budowie modeli scoringowych.

W rozdziale trzecim zostały zaprezentowane metody strojenia zapobiegające nadmiernemu dopasowaniu modelu do danych oraz metody oceny skuteczności klasyfikacji modeli.

Rozdział czwarty przedstawia wyniki analizy empirycznej, w której dokonano porównania skuteczności siedmiu zastosowanych metod klasyfikacji, służących do budowy modeli scoringowych. W szczególności, dla modelu regresji logistycznej zbudowano kartę scoringową oraz wyznaczono optymalny punkt odcięcia. Z kolei, dla najbardziej skutecznego z modeli data mining wyznaczono najistotniejsze cechy kredytobiorcy oraz określono siłę i kierunek ich wpływu na zmienną prognozowaną.

Rozdział 1. Charakterystyka credit scoringu w finansach

1.1 Ryzyko kredytowe w kontekście działalności bankowej

1.1.1 Istota ryzyka w działalności banku

Ryzyko jest zjawiskiem, które towarzyszy ludziom od zarania dziejów. Związane jest ono z podejmowaniem decyzji, których rezultat jest przewidywalny w bliższej lub dalszej przyszłości. Nie można go uniknąć, gdyż wiąże się z niepewnością, której nie da się kontrolować ani precyzyjnie przewidzieć. W każdej więc chwili jesteśmy narażeni na jego obecność. Szczególne miejsce ryzyko zajmuje w prowadzeniu działalności gospodarczej, gdzie podmioty gospodarcze mają z nim do czynienia każdego dnia. W naukach ekonomicznych przyjmuje ono rangę kategorii pierwszoplanowej[1].

Działalność banku jest także nierozłącznie związana z występowaniem ryzyka. Można by zaryzykować stwierdzenie, że jedną z podstawowych działalności banku jest identyfikacja różnych typów ryzyka, ocena jego wielkości oraz wprowadzanie odpowiednich procedur zarządzania ryzykiem. Mając na uwadze specyfikę działalności banków trzeba wspomnieć o ryzyku związanym z poziomem stóp procentowych, płynnością finansową, działalnością kredytową, czy też ryzykiem inwestycyjnym występującym na rynku kapitałowym. Należy podkreślić, że konieczność odpowiedniego podejścia do zjawiska ryzyka nie jest wyłącznie sprawą danego banku. Za zapewnienie bezpiecznego utrzymania powierzonych środków finansowych oraz gwarancję odpowiednich standardów w zakresie ryzyka jest również odpowiedzialny bank centralny. Fundamentalną rolę pełni też tutaj odpowiednie prawodawstwo[2].

Jednym z najważniejszych rodzajów ryzyka występującym w działalności banków jest ryzyko kredytowe, określane często mianem typowego ryzyka bankowego, inaczej też nazywane ryzykiem podmiotowym. Należy ono raczej do kategorii ryzyka czystego, tj. odwołującego się do poniesienia straty[3]. Generalnie jest to ryzyko związane z tym, że klient banku nie będzie w stanie zagwarantować wymaganych środków pieniężnych do realizacji transakcji, zwykle z powodu problemów z płynnością finansową lub bankructwa[4].

[1] M. Wójciak, *Metody oceny ryzyka kredytowego,* Polskie Wydawnictwo Ekonomiczne, Warszawa 2007, s. 11.
[2] J. Wątroba, *Skoring kredytowy a modele data mining*, StatSoft, 2004, s. 65-66.
[3] M. Wójciak, op. cit., s. 13.
[4] O. Rusak, *Ryzyko kredytowe jako jedno z ryzyk w działalności banku*, Zeszyty Naukowe Uniwersytetu Przyrodniczo-humanistycznego w Siedlcach, 2015, nr 104, s. 269-270.

1.1.2 Rodzaje ryzyka kredytowego

Na podstawie literatury przedmiotu można określić kilka różnych kryteriów podziału ryzyka kredytowego. W zależności od elementów bilansu banku go tworzących, ryzyko kredytowe najczęściej dzieli się na[5]:

1) **ryzyko aktywne** – zwane również czynnym, rozumiane jako zagrożenie związane z brakiem spłaty zadłużenia przez kredytobiorcę zgodnie z terminem określonym w umowie kredytowej, co w następstwie może skutkować utraceniem płynności przez bank. Cechą tego rodzaju ryzyka jest to, że w zdecydowanej mierze jest ono zdeterminowane przez bank, z tym że pewna część tego ryzyka ma charakter, który jest od niego niezależny, co powoduje, że nie można go w pełni kontrolować;
2) **ryzyko pasywne** – nazywane także biernym, odnosi się do wcześniejszego niż przewiduje umowa kredytowa, wycofania środków finansowych przez klienta (mogą nim być np. przedsiębiorstwa, jednostki samorządowe czy gospodarstwa domowe), a także groźby nieotrzymania kredytu refinansowego od innych instytucji finansowych. W takich okolicznościach bank jest stroną bierną, która nie ma zbyt dużego wpływu na tego typu ryzyko.

Z punktu widzenia pierwotnych źródeł ryzyko kredytowe można podzielić na[6]:

1) **ryzyko egzogeniczne** (zewnętrzne) – dotyczy czynników egzogenicznych, do których zalicza się m.in. cechy klienta banku (np. skłonność lub awersję do spłaty kredytu), czynniki makroekonomiczne (poziom bezrobocia, sytuacja gospodarcza, stopa inflacji, sytuacja polityczna czy społeczna), a także siły natury. Wymienione czynniki mają istotny wpływ na spłatę zaciągniętych zobowiązań finansowych kredytobiorcy w przeciwieństwie do banku, który nie ma na nie praktycznie żadnego wpływu;
2) **ryzyko endogeniczne** (wewnętrzne) – to rodzaj ryzyka o charakterze subiektywnym, dotyczącym istoty działania instytucji finansowej. Można powiedzieć, że ryzyko endogeniczne zależy przeważnie od sprawności systemów, które odpowiadają za podejmowanie decyzji kredytowych. Mówimy tutaj o przyjętych mechanizmach oceny danego klienta i metodach monitorowania ryzyka kredytowego, a także o sposobach podziału kompetencji kredytowych wśród pracowników banku i określeniu zakresu ich odpowiedzialności.

[5] A. Kuchciński, *Ryzyko kredytowe w działalności* banku, Kwartalnik Naukowy Uczelni Vistula, 2016, nr 2(48), s. 38-39.
[6] Ibidem, s. 40.

1.1.3 Zarządzanie ryzykiem kredytowym

Trudności z działalnością kredytową mają istotny wpływ na kondycję banku. Zakłócenia w spłacie zobowiązań finansowych wobec banku powodują bowiem zmiany prognozowanych wpływów środków pieniężnych. W sytuacji poważnych problemów, skutkują one koniecznością utworzenia rezerw celowych na kredyt i restrukturyzację przyszłych spłat kredytu, a w skrajnym przypadku prowadzą do spisania kredytów w straty. Wysoki odsetek złych kredytów pogarsza płynność banków oraz może prowadzić nawet do upadłości. Z tego właśnie powodu zarządzanie ryzykiem kredytowym zaczyna odgrywać coraz większą rolę w procesie zarządzania bankiem[7].

W związku z tym kompletna analiza ryzyka kredytowego powinna zawierać wszystkie aspekty mające wpływ na wysokość tego ryzyka. W praktyce bankowej wymienia się pięć głównych kryteriów, tzw. Pięć C kredytu. Jest to najbardziej popularny system ekspercki, który ma służyć prawidłowości i kompletności procesu kredytowania. Jego nazwa powstała od pierwszych liter pięciu czynników, a są to[8]:

1) **charakter** (*Character*) – to najważniejsza cecha budująca wizerunek kredytobiorcy (ocena reputacji). Chodzi tutaj o takie cechy jak: uczciwość, stabilność czy solidność. Najłatwiej jest ją ocenić w przypadku, gdy kredytobiorca jest stałym klientem banku.;
2) **kapitał** (*Capital*) – posiadany kapitał;
3) **zdolność kredytowa** (*Capacity*) – to możliwość spłaty zaciągniętego kredytu wraz z odsetkami w terminach ustalonych w umowie;
4) **zabezpieczenie** (*Collateral*) – to wtórne źródło uzyskania należności, a więc w sytuacji upadłości kredytobiorcy może ono rekompensować straty poniesione z tytułu niespłaconego kredytu;
5) **uwarunkowania zewnętrzne** (*Conditions*) – to istotny element, który należy mieć na uwadze, gdyż w fazie regresji cyklu koniunkturalnego wzrasta prawdopodobieństwo niespłacenia kredytu.

Ocena zdolności kredytowej jest podstawowym składnikiem systemu zarządzania indywidualnym ryzykiem kredytowym[9]. W toku swej bogatej praktyki banki wypracowały

[7] A. Matuszczyk, *Credit scoring,* Wydawnictwo CeDeWu, Warszawa 2018, s. 31.
[8] M. Wójciak, op. cit., s.16.
[9] A. Matuszczyk, *Ryzyko kredytowe*, w: Zarzadzanie ryzykiem w banku komercyjnym, pod red. M. Iwanicz-Drozdowksa, Wydawnictwo Poltext, Warszawa 2017, s. 135-136.

rozmaite metody oceny zdolności kredytowej. Można tu wydzielić dwie podstawowe kategorie zdolności kredytowej[10]:

- **zdolność kredytową pod względem formalnoprawnym** – wiarygodność prawna kredytobiorcy,
- **zdolność kredytową pod względem merytorycznym** – wiarygodność ekonomiczna kredytobiorcy.

Przez wiarygodność prawną kredytobiorcy rozumiemy zdolność do podejmowania czynności prawnych, a więc zdolność prawną do zawarcia umowy z bankiem. Z kolei wiarygodność ekonomiczna kredytobiorcy jest terminem bardziej złożonym i zawiera dwa podstawowe aspekty oceny: ekonomiczny i personalny. Z personalnego punktu widzenia rozpatruje się elementy determinujące zaufanie do kredytobiorcy, do których zalicza się: dotychczasowe doświadczenie, zdolności menedżerskie, reputację oraz ocenę etyczną, czyli solidność, odpowiedzialność i lojalność kredytobiorcy za interesy prowadzonej firmy[11].

Ekonomiczne aspekty oceny merytorycznej zdolności kredytowej w zasadzie sprowadzają się do oceny elementów opisujących dotychczasową i perspektywiczną sytuację ekonomiczno-finansową danego kredytobiorcy oraz jakość zabezpieczeń prawnych kredytu[12].

W zależności od tego, czy kredyt dotyczy osoby prawnej czy osoby fizycznej, badanie zdolności kredytowej ma specyficzne cechy. W sytuacji kredytów gospodarczych, dominującą rolę w ocenie zdolności kredytowej odgrywa aspekt ekonomiczny. W odniesieniu natomiast do osób fizycznych i przyznawanych im kredytów konsumpcyjnych, większą wagę ma indywidualna analiza osoby kredytobiorcy. Trzeba zauważyć, że w metodyce badania zdolności kredytowej nie ma jednolitego wzorca[13].

Podstawowymi źródłami informacji o kredytobiorcy wykorzystywanymi do oceny zdolności kredytowej są[14]:

[10] E. Świątczak, *Badanie zdolności kredytowej przedsiębiorstwa jako sposób na ograniczenie ryzyka kredytowego banku w procesie kredytowania przedsiębiorstw*, Zeszyty Studenckie Wydziału Ekonomicznego Uniwersytetu Gdańskiego „Nasze Studia", 2009, nr 4, s. 61.
[11] M. S. Wiatr, R. Jagiełło, Ryzyko kredytowe, [w:] Współczesna bankowość, red. M. Zaleska, Difin, Warszawa 2007, s. 78.
[12] Ibidem., s. 309.
[13] A. Matuszczyk, *Ryzyko kredytowe*, op. cit., s. 138.
[14] A. Matuszczyk, *Credit scoring,* op. cit., s.33.

- **historia dotychczasowej współpracy z bankiem** – będące w posiadaniu rachunki bankowe, wykorzystywane produkty bankowe, reklamacje i skargi, lojalność wobec banku;
- **informacje we wnioskach kredytowych** – dane osobiste, wiek, stan rodziny, własność majątkowa, liczba osób na utrzymaniu, sytuacja zarobkowo-dochodowa, okres i stabilność zatrudnienia, charakter profesji;
- **opinie i referencje z innych banków** – rodzaje kredytów bankowych, stany oraz przebieg obsługi zadłużenia, rodzaje posiadanych rachunków, terminy ich zamknięcia i otwarcia, przeciętne stany sald,
- **biura informacji kredytowej** – wyodrębnione agencje dostarczające informacje o dotychczasowej wiarygodności kredytowej klientów z różnych banków, instytucji finansowych i pozafinansowych,

Uzyskane z różnych źródeł informacje są przetwarzane i odpowiednio wartościowane, tak, aby mogły służyć ocenie.

W literaturze opisywane są różne podejścia stosowane przy badaniu zdolności kredytowej. Za najpopularniejszą metodę oceny ryzyka kredytowego można uznać scoring kredytowy. Metoda ta sprowadza się do wyznaczenia na podstawie różnych charakterystyk kredytobiorcy punktowej oceny, która następnie stosowana jest do klasyfikowania kredytobiorcy do grupy o zdefiniowanym poziomie ryzyka[15].

1.2 Credit scoring – charakterystyka ogólna

Podstawowa koncepcja metody credit scoring jest oparta na założeniu, iż ryzyko kredytowe jest mierzalne i można je mierzyć, wykorzystując do tego dane na temat osobistych cech klienta ubiegającego się o kredyt. Zawarte we wniosku kredytowym oraz pozostałych załącznikach dane osobiste i informacje o sytuacji majątkowej aplikującego, zostają skwantyfikowane, tj. przekształcone do wartości cyfrowych, co pozwala identyfikować rodzaj i poziom ryzyka kredytowego, usprawniając w ten sposób proces oceny zdolności kredytowej oraz ułatwiając podjęcie decyzji przyznania kredytu[16].

[15] G. Migut, J. Wątroba, *Skoring kredytowy a modele data mining*, Rynek terminowy, nr 1, 2015, s. 48.
[16] D. Prokopowicz, *Główne determinanty zastosowania metody credit scoring w zarządzaniu ryzykiem kredytowym*, Kwartalnik Naukowy Uczelni Vistula, 2015, nr 1(43), s. 19.

1.2.1 Pojęcie credit scoringu

W literaturze przedmiotu można znaleźć wiele definicji credit scoringu. Pierwsza interpretacja określenia credit scoringu, którą warto przedstawić zawarta jest w amerykańskiej Ustawie o Równości Szans w Udzielaniu Kredytu z 1975 roku. Zgodnie z tym aktem prawnym, system credit scoringu objaśnia się jako: „wprowadzany w wyniku doświadczenia, uzasadnionego statystykami i dowodami oraz:

- oparty na danych pochodzących z empirycznego porównania przykładowych populacji lub grup rzetelnych i nierzetelnych kandydatów, którzy aplikowali o kredyt w określonym przedziale czasowym,
- zbudowany w celu oceny zdolności kredytowej klientów przy respektowaniu interesu kredytodawcy,
- uzasadniony i rozwijany poprzez zastosowanie zaakceptowanych statystycznych zasad oraz metodologii,
- periodycznie weryfikowany poprzez użycie odpowiednich metodologii i zasad statystycznych oraz modyfikowany w celu zapewnienia zdolności przewidywania”[17].

Credit scoring definiowany jest również jako system używany przez kredytodawców z zamiarem określenia, czy należy udzielić kredytu danemu klientowi. Informacje, takie jak: liczba i rodzaj posiadanych rachunków bankowych, historia płacenia rachunków, czas posiadania rachunku, nieuregulowane zadłużenie, czy opóźnienia w płatnościach są pobierane z wniosku kredytowego i raportu kredytowego. Wykorzystując model statystyczny, kredytodawcy porównują zebrane informacje z historią kredytową klientów o podobnych cechach. Następnie system credit scoring przyznaje punkty czynnikom, które zwiększają prawdopodobieństwo spłacenia kredytu. Całkowita suma uzyskanych punktów pozwala przewidzieć, czy kredyt zostanie spłacony i płatności będą regulowane terminowo[18].

Inne podejście określa credit scoring jako termin używany do opisania tradycyjnych metod statystycznych stosowanych do klasyfikacji kredytobiorców na grupy o złym i dobrym ryzyku kredytowym. Definiowanie credit scoringu jako metody binarnej klasyfikacji

[17] Legal Information Institute, https://www.law.cornell.edu/cfr/text/12/202.2 (dostęp: 15.10.2019).
[18] „RHOL Credit Scoring” http://www.rental-housing.com/fcra/PDF/scoring.pdf (dostęp: 15.10.2019).

kredytobiorców stało się bardzo popularne z powodu gwałtownego rozwoju rynku kredytów konsumpcyjnych[19].

Jeszcze inna definicja stanowi, że credit scoring jest sposobem modelowania statystycznego w ramach reprezentatywnej bazy danych kredytobiorców i uzyskania wyniku liczbowego dla każdego kredytu. Powstały wynik kredytowy jest liczbą, w oparciu o którą można przewidzieć zdolność kredytobiorcy do spłacenia zadłużenia. Wynik ten może być stosowany w klasyfikacji pojedynczych kredytów do właściwych grup ryzyka. Dysponując całą bazą wyników, banki ustalają punkt odcięcia (*cut-off point*), na podstawie tego jak duże ryzyko kredytowe są w stanie zaakceptować[20].

Od strony technicznej ilościowa ocena ryzyka kredytowego polega na wyliczeniu prognozy dla kategorycznej zmiennej o rozkładzie dwumianowym (zmienna objaśniana). Wartości zmiennej zależnej oznaczają przynależność osoby ubiegającej się o kredyt do jednej z dwóch grup: osoby, którym nie powinno się przyznawać kredytu (z uwagi na duże ryzyko dla banku) oraz osoby, którym powinno się przyznać kredyt (małe ryzyko). Lista potencjalnych zmiennych niezależnych (predyktorów) obejmuje różnego typu informacje ilościowe i jakościowe zarówno na temat samego kredytobiorcy, jak i jego otoczenia. W praktyce ilościowa ocena ryzyka oznacza wartość liczbową z określonego przedziału (zazwyczaj od 0 do 100 punktów). Uzyskana wartość prognozy wyliczona dla konkretnego klienta stanowi szacunkowy poziom ryzyka wiarygodności kredytowej danego wniosku kredytowego[21].

Ocena punktowa powstaje na podstawie utworzonego wcześniej modelu scoringowego. Kształt takiego modelu ustala się w oparciu o doświadczenia banku z osobami, które wcześniej ubiegały się o przyznanie kredytu. Takie podejście zakłada, że dany kredytobiorca będzie wykazywał podobne zachowania do historycznych kredytobiorców o zbliżonych cechach do niego. Zadaniem analityka korzystającego z tego typu metod jest adekwatny dobór zmiennych opisujących zachowanie osoby starającej się o kredyt i zbudowanie na ich podstawie modelu, który byłby w stanie rozpoznać, czy dany klient jest wiarygodny, czy też nie. Taki model powinien mieć zdolność do uogólnienia informacji

[19] D.J. Hand, W.E Henley, *Statistical Classification Methods in Consumer Credit Scoring: a Review*, „Journal of the Royal Statistical Society", 1997, Part 3, s. 522.
[20] T.H. Stanton, *Credit Scoring and Loan Scoring: Tools for Improved Management of Federal Credit Programs*, Grant Report, July 1999, s. 8.
[21] J. Wątroba, *Skoring kredytowy a modele data mining*, Statsoft Polska, 2015, s. 67.

ujętych w danych historycznych i wykazywać się podobną skutecznością również dla nowych, nieznanych wcześniej modelowi danych[22].

Można zatem stwierdzić, że credit scoring jest metodą oceny ryzyka kredytowego związanego z danym kredytobiorcą. Jest to statystyczno-matematyczny instrument umożliwiający bankowi przyporządkowanie klienta do odpowiedniej kategorii ryzyka.

1.2.2 Geneza credit scoringu

Początków credit scoringu można doszukiwać się w sformułowanej w 1936 roku przez Ronalda Fishera metodzie polegającej na klasyfikacji różnych grup w populacji na podstawie kojarzenia cech pokrewnych[23]. Przedmiotem badań brytyjskiego statystyka były irysy i możliwość ich identyfikacji dzięki pomiarom na tle pozostałych roślin. Wykorzystując założenia z badań Fishera, w 1941 roku David Durand wskazał, że tę metodę statystyczną można także zastosować w analogiczny sposób do odróżnienia złych od dobrych kredytów[24]. Jego czysto teoretyczne badania prowadzone w ramach projektu badawczego dla National Bureau of Economic Research US nie zostały jednak zastosowane w praktyce[25].

Jednocześnie w latach 30. XX w. niektóre firmy zajmujące się sprzedażą wysyłkową wprowadziły liczbowe systemy scoringowe, aby przełamać niekonsekwencje w podejmowaniu decyzji pożyczkowych przez analityków kredytowych. Niestety wybuch II wojny światowej spowodował, że analitycy kredytowi zaczęli być masowo powoływani do wojska, co doprowadziło do powstania poważnego niedoboru ludzi specjalizujących się w tej dziedzinie finansów. Wobec tego wszystkie firmy wysyłkowe i instytucje finansowe zaczęły borykać się z problemami zarządzania ryzykiem kredytowym. W związku z tym, aby zmniejszyć ryzyko niespłacenia kredytu, firmy zlecały analitykom opracowanie zasad, które można by było stosować przy podejmowaniu decyzji kredytowej. Te reguły były następnie stosowane przez pracowników mniej wykwalifikowanych w celu wsparcia podejmowania decyzji kredytowych[26].

[22] Ibidem, s. 67-68.
[23] R.A. Fisher, *The use of multiple measurements in taxonomic problems*, „Annual Eugenics", 1936, nr 7, s.179 – 188.
[24] D. Durand, *Risk Elements in Consumer Instatement Financing,* National Bureau of Economic Research, New York, 1941.
[25] A. Janc, M. Kraska, *Credit-scoring. Nowoczesna metoda oceny zdolności kredytowej*, Bibloteka Menadżera i Bankowca, Warszawa 2001, s. 10.
[26] A. Matuszczyk, *Credit scoring,* op. cit., s.50.

Po II wojnie światowej wprowadzono jeszcze kilka innych metodologii w zakresie oceny ryzyka kredytowego, jednak niemających większego wpływu na rozwój credit scoringu. Dopiero po powstaniu pierwszej firmy konsultacyjnej – *Fair Isaac and Company*, założonej w 1956 roku przez matematyka Earla Isaaca i inżyniera Williama Faira w San Francisco, datuje się znaczący rozkwit metody scoringowej. W literaturze bardzo często właśnie ich przywołuje się jako głównych autorów sukcesu credit scoringu[27].

Następnym punktem przełomowym w rozwoju credit scoringu było pojawienie się na rynku bankowym kart kredytowych w późnych latach sześćdziesiątych, co skłoniło banki oraz emitentów tych kart do uświadomienia sobie przydatności i praktyczności credit scoringu. Dynamicznie wzrastająca liczba klientów ubiegających się o przyznanie kart kredytowych sprawiła, że instytucje finansowe zmuszone zostały do zautomatyzowania procesu podejmowania decyzji pożyczkowych, zarówno ze względów ekonomicznych, jak i z powodu braku pracowników. Po pewnym czasie od wprowadzenia systemu credit scoring instytucje pożyczkowe zdały sobie sprawę, że system ten jest o wiele bardziej skuteczny, niż jakikolwiek wcześniejszy system oceny zdolności kredytowej, a współczynnik niespłaconych kredytów spadł o 50%[28].

Kolejnym kluczowym wydarzeniem w popularyzacji credit scoringu, zapewniającym jego całkowitą akceptację było wprowadzenie w życie w Stanach Zjednoczonych ustawy o Równości Szans w Udzielaniu Kredytu (ECOA – *Equal Credit Opurtunity Act*) w 1975 roku. Ten akt prawny zakazywał jakiejkolwiek dyskryminacji wobec kredytobiorcy, chyba że dyskryminacja byłaby uzasadniona statystycznie w ramach oceny zdolności kredytowej[29].

W latach osiemdziesiątych sukces credit scoringu przy zakładaniu kart kredytowych spowodował, że banki zaczęły stosować system punktowej oceny ryzyka również dla innych produktów finansowych, takich jak pożyczki osobiste, a w następnym okresie, także przy udzielaniu kredytów gospodarczych i hipotecznych. Z kolei w latach dziewięćdziesiątych postęp technologiczny umożliwił wypróbowanie innych metod do budowy kart punktowych. W obliczu gwałtowanego rozwoju bezpośredniego marketingu, większe przedsiębiorstwa

[27] A. Janc, M. Kraska, op. cit., s. 11-12.
[28] L.C. Thomas, *A Survey of Credit and Behavioural Scoring: Forecasting financial risk of lending to consumers*, University of Edinburgh, UK, Credit Research Centre Working Papers, No. 99/2, s.6-7.
[29] Ibidem, s. 7.

zaczęły stosować system kart punktowych także w celu poprawienia wskaźnika odpowiedzi na kampanie reklamowe.[30].

Z nastaniem epoki komputerów systemy ocen punktowych stały się zaawansowanymi modelami prognostycznymi, które na początku bazowały głównie na modelu regresji liniowej i logistycznej. Obecnie śmiało można to pojęcie rozwinąć na wiele innych metod modelowania predykcyjnego, wliczając w to techniki Data Mining: drzewa decyzyjne, lasy losowe, sieci neuronowe, czy też wiele innych metod ciągle się rozwijających. Nie powinno się też credit scoringu identyfikować tylko z bankowym procesem oceny ryzyka. Stosuje się go dziś również w szeregu innych procesach, w których klient zawierający umowę, najczęściej zobowiązujący się do periodycznych obciążeń finansowych (np. abonament telefoniczny), musi być przed zawarciem umowy wstępnie oceniony, by instytucja dostarczająca dane usługi nie naraziła się na straty związane z brakiem terminowych spłat zobowiązań[31].

Obecnie w kontekście technologii Big Data mówi się o nowej erze, a wszelkie analizy scoringowe są doskonałym tego przykładem, zwłaszcza stosowanym przy bardzo prostym modelowaniu biznesowym. Właśnie w credit scoringu ukształtowały się wszystkie pożądane składowe modelowania predykcyjnego, takie jak: dobór próby, proste modele biznesowe, rozumienie populacji, testowanie na różnych próbach, walidacja modeli, ocena modeli, analiza wpływu wniosków odrzuconych, testowanie strategii, wyznaczenie punktów odcięcia, kalibracja do wartości prawdopodobieństwa, wdrożenie w systemie decyzyjnym oraz testowanie po zaimplementowaniu i monitoring. Cały cykl życia modelu można zaobserwować na przykładzie credit scoringu[32].

Wzrost znaczenia ilościowego podejścia do zarządzania ryzykiem wiązało się z coraz bardziej złożonymi produktami finansowymi, które trudno było analizować w sposób opisowy. Innowacje w zakresie oceny ryzyka kredytowego były rezultatem kilku sił, w tym[33]:

- rozwinięcia rynków kredytowych w celu zagospodarowania nowych sektorów kredytowych, zarówno na rynku międzynarodowym, jak i lokalnym;

[30] A. Matuszczyk, *Credit scoring, op. cit.,* s. 51.
[31] K. Przanowski, *Credit scoring. Studia przypadków procesów* biznesowych, Oficyna Wydawnicza Szkoła Główna Handlowa w Warszawie, Warszawa 2015, s. 25.
[32] Ibidem, s.25-26.
[33] A. Matuszczyk, *Ryzyko kredytowe*, op. cit., s. 147.

- deregulacji, które stymulowały wprowadzenie innowacji finansowych i umożliwiły kolejnym graczom świadczenie usług;
- wzrostu ryzyka operacji pozabilansowych;
- sekurytyzacji, która pobudziła rozwój bardziej efektywnych (i standardowych) narzędzi służących wydajnemu pomniejszaniu ryzyka kredytowego;
- ograniczenia marż na pożyczkach, co zobligowało banki do poszukiwania mniej kosztownych metod pomiaru i zarządzania ryzykiem kredytowym;
- rozwoju teorii finansów, który dostarczył innego podejścia do ryzyka kredytowego;
- reform regulacyjnych, w tym Solvency dla firm ubezpieczeniowych i Bazylei 2 dla banków;
- rozwoju pozagiełdowego obrotu derywatami kredytowymi.

1.2.3 Klasyfikacja credit scoringu

Punktową ocenę ryzyka kredytowego można klasyfikować według następujących kryteriów[34]:

1. ze względu na cel analizy skoringowej:
   a) systemy minimalizujące ryzyko niewywiązania się klienta ze zobowiązań,
   b) systemy maksymalizujące zysk wypracowany w związku ze współpracą z danym klientem;
2. ze względu na oceniany podmiot:
   a) skoring podmiotów gospodarczych,
      – małych przedsiębiorstw,
      – średnich i dużych przedsiębiorstw,

   b) skoring osób fizycznych,

3. ze względu na rodzaj kredytu:
   a) skoring kredytów gospodarczych,
   b) skoring kredytów konsumpcyjnych,
      – kredytów samochodowych,
      – kredytów hipotecznych
      – kredytów gotówkowych,

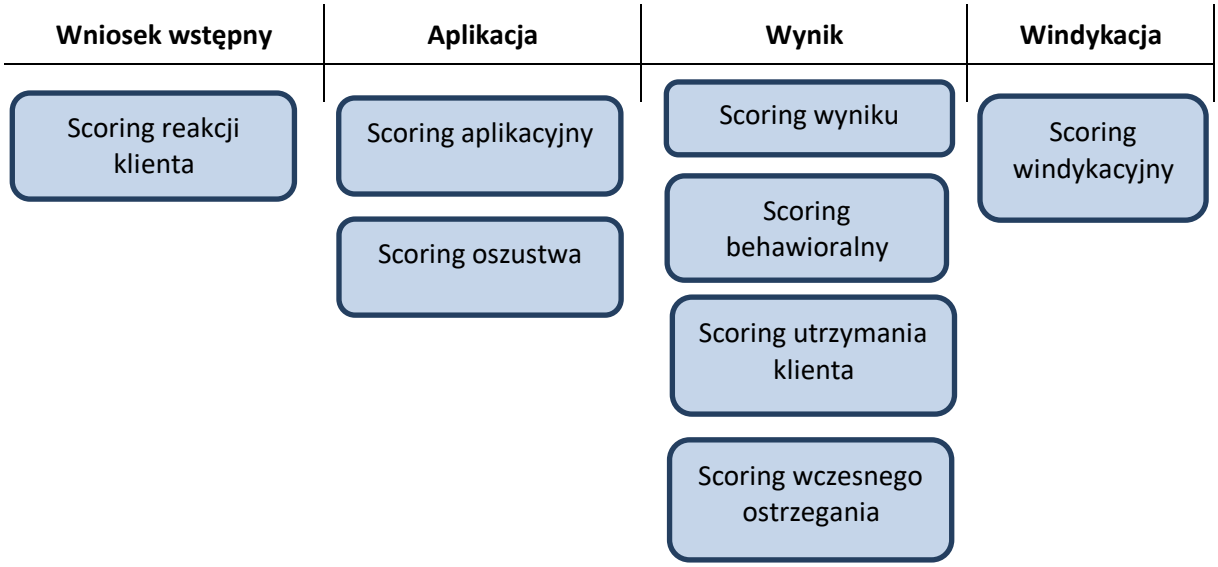
[34] A. Janc, M. Kraska, op. cit., s. 35-36.

– wykorzystywany w procesie wydawania kart kredytowych,

4. ze względu na cel wykorzystania:
   a) wewnętrzny,
   b) zewnętrzny;
5. ze względu na podmiot dokonujący punktowej oceny ryzyka kredytowego:
   a) ocena punktowa biura kredytowego,
   b) ocena punktowa banku centralnego,
   c) ocena emitentów kart płatniczych,
   d) ocena towarzystwa ubezpieczeniowego,
   e) ocena punktowa agencji ratingowych,
   f) ocena punktowa banku komercyjnego,
   g) inne;
6. ze względu na dziedzinę, w której skoring jest wykorzystywany:
   a) skoring bankowy,
   b) skoring marketingowy,
   c) skoring ubezpieczeniowy,
   d) inne (wykorzystywane między innymi w kwestiach podatków, medycynie itd.).

W różnych etapach życia kredytu można zastosować różne rodzaje scoringu. Poniżej zaprezentowano i opisano wybrane rodzaje scoringu, istotne z perspektywy działalności kredytowej.

Rys. 1.1. Różne rodzaje scoringu na poszczególnych etapach życia kredytu



Źródło: A. Matuszczyk, *Ryzyko kredytowe*, w: Zarzadzanie ryzykiem w banku komercyjnym, pod red. M. Iwanicz-Drozdowksa, Wydawnictwo Poltext, Warszawa 2017, s. 155.

Skoring reakcji klienta (*response scoring*) określa prawdopodobieństwo reakcji klienta na kampanię marketingową. Banki starają się spersonalizować ofertę produktu finansowego pod danego klienta i przewidzieć, czy klient dokona zakupu produktu. Skoring ten ma na celu obniżenie kosztów pozyskania klientów oraz zminimalizowanie jego niezadowolenia i niedogodności związanego z ofertą banku[35].

Scoring aplikacyjny (*application score*) określany również jako scoring użytkowy, stosowany jest do oceny wniosku o kredyt lub o inny produkt finansowy. Ma on na celu obliczenie, jakie jest prawdopodobieństwo, że kredytobiorca wywiąże się ze swoich zaciągniętych zobowiązań. W skrócie polega to na analizie złożonych przez kredytobiorcę odpowiedzi na zestaw pytań zawartych we wniosku o kredyt. Stosowna liczba punktów jest kojarzona z konkretną odpowiedzią. Następnie uzyskane punkty są sumowane i jeśli zdefiniowany przez bank próg (tzw. punkt odcięcia – *cut off point*) zostanie przekroczony, wtedy kredyt jest przyznawany; w innym wypadku wniosek zostaje odrzucony[36].

Scoring oszustwa (*fraud scoring*), tworzony jest w celu ograniczenia strat powstałych na skutek wyłudzeń. Modele te służą do określenia prawdopodobieństwa, czy mamy do czynienia z wyłudzeniem, zabezpieczając w ten sposób przed uruchomieniem niewłaściwej transakcji. W modelach tych następuje automatyczna weryfikacja z zewnętrznymi i wewnętrznymi bazami danych, przeprowadza się porównanie z wcześniej składanymi wnioskami oraz następuje wyszukanie niespójności, tj. nietypowych, nieprawidłowych lub skrajnych wartości. W dalszej kolejności pozyskane dane są odpowiednio oceniane i w rezultacie tej oceny wniosek może zostać oddalony, bądź wysyłane jest ostrzeżenie i wniosek poddawany jest dodatkowo szczegółowej analizie[37].

Scoring behawioralny (*behavioural scoring*) jest kolejną techniką oceniającą, która umożliwia kredytodawcy podjąć właściwą decyzję dotyczącą stałego klienta banku przez przewidywanie wyników jego działalności. Podstawową zatem różnicą pomiędzy scoringiem behawioralnym a scoringiem użytkowym (aplikacyjnym) jest fakt, że scoring użytkowy dotyczy nowych klientów banku, natomiast ocena behawioralna odnosi się do stałych

[35] T. Gestel, B. Baesens, *Credit Risk Management. Basic Concepts: financial risk components, rating analysis, models, economic and regulatory capital,* Oxford University Press, New York 2009, s. 96-97.
[36] A. Matuszczyk, *Ryzyko kredytowe*, op. cit., s. 155.
[37] G. Ignaciuk, *Zastosowanie metod scoringowych w działalności bankowej,* Statsoft, 2010, s. 24.

klientów. Zasady matematyczne, w oparciu o które budowana jest tablica scoringowa, są takie same dla obu rodzajów technik[38].

Scoring behawioralny odpowiada na pytanie, czy dany klient byłby chętny na dodatkowy produkt z oferty. Patrząc od strony banku oznacza to na przykład, czy można klientowi zaoferować kolejny kredyt, przedłużenie umowy na kartę kredytową, czy zwiększenie limitu w karcie. Do oceny punktowej behawioralnej stosowane są dane z możliwych dostępnych źródeł: wcześniejszego scoringu użytkowego, innego scoringu behawioralnego (jeśli taki miał miejsce), oraz informacje uzyskane z obserwacji klienta dokonywanej periodycznie, na przykład obserwacja obrotów na koncie bankowym, średni poziom zadłużenia lub terminowość spłat kredytu[39].

Ocenę behawioralną uznaje się za bardziej wiarygodną i bardziej kompletną, gdyż stosuje się w niej dane historyczne zgromadzone przez bank w trakcie współpracy z klientem, a nie wyłącznie dane z wniosku kredytowego. Ponadto z uwagi na to, że ocena ta dokonywana jest okresowo, jest ona regularnie uaktualniana. Stanowi jednak metodę bardziej skomplikowaną, jeśli chodzi o dostosowanie do potrzeb i implementację[40].

Scoring zysku (*profit scoring*) należy do stosunkowo nowych koncepcji wykorzystujących metodę scoringu. Idea tej techniki powstała z przekonania, że to zysk, a nie wskaźnik określający ryzyko braku spłaty zadłużenia, jest bardziej użytecznym kryterium w procesie podejmowania decyzji związanej z przyznaniem kredytu. Z tego powodu właśnie powstały modele profit scoring, które są rozwinięciem podstawowych modeli scoringowych. W modelach tych uwzględnia się zarówno prawdopodobieństwo spłacenia kredytu przez kredytobiorcę, jak i związany z tym kredytem zysk. Jest to narzędzie bardziej zaawansowane, bowiem zawiera szereg dodatkowych czynników ekonomicznych, takich jak polityka cenowa, marketing czy poziom obsługi. Implementacja modelu, jakim jest scoring zysku, jest związana z wprowadzeniem do metodologii dodatkowych elementów, do których wymagane są odpowiednie bazy danych, mianowicie potrzebne są zintegrowane systemy informatyczne przechowujące i przetwarzające kluczowe informacje o klientach oraz dobór właściwej techniki modelowania wykorzystywanej w ocenie przyszłego zysku. Profit scoring jest wciąż

[38] A. Janc, M. Kraska, op. cit., s. 37.
[39] A. Staniszewska, *Zarządzanie portfelem kredytowym banku*, Oficyna Wydawnicza Szkoła Główna Handlowa w Warszawie, Warszawa 2012, s. 163.
[40] A. Matuszczyk, *Credit scoring,* op. cit., s. 70.

nową techniką, systematycznie rozwijaną, ciągle jednak brakuje modeli na tyle efektywnych, aby były w stanie bez większych problemów być wykorzystywane[41].

Scoring wczesnego ostrzegania (*early warning scoring*). Zadaniem tego systemu jest detekcja potencjalnych ryzyk związanych z partnerem transakcji. Wyszukani przez model partnerzy są umieszczani na listę obserwacyjną w celu szczegółowej inspekcji. Systemy te mogą być też postrzegane jako specyficzne przypadki scoringu wyniku o wąskim horyzoncie czasowym, tj. w okresie od 6 do 12 miesięcy. Wykorzystywane są w nich dane makroekonomiczne, informacje o kapitale, cenach obligacji skarbowych oraz derywatach[42].

Scoring windykacyjny jest narzędziem wspierającym proces podejmowania decyzji dotyczących zarządzania nieściągalnym długiem. Umożliwia estymacje poziomów odzysku, które są prognozowane do uzyskania z określonych zobowiązań klientów. Scoring windykacyjny wspomaga działania mające na celu zwiększenie odzysków, usprawnia proces zarządzania nieobsługiwanymi zobowiązaniami, przez co powoduje poprawę efektywności restrukturyzacji i windykacji, umożliwia efektywniejsze wykorzystanie zasobów osobowych oraz zmniejsza koszty odpisów z tytułu utraty wartości [43].

Scoring utrzymania klienta ma na celu zmniejszenie grupy odchodzących klientów. Jest wykorzystywany do identyfikowania, który klient najprawdopodobniej zamknie konto bankowe lub znacząco zredukuje swoją aktywność. Scoring ten jest ważnym wsparciem dla systemu zarządzania relacjami z klientem[44].

Alternatywą, czy raczej – dopełnieniem systemów credit scoringu w ocenie wiarygodności kredytowej firm – jest specyficzna grupa modeli etykietowanych wspólną nazwą Z-score[45].

Podstawową różnicą pomiędzy modelami Z-score a credit scoringiem jest przedział czasu, jakiego dotyczy analiza zdolności kredytowej. Główną różnicą, na którą trzeba wskazać, jest uwzględnianie w modelach Z-score wyłącznie czynników ilościowych. Podstawą oceny są przeważnie odpowiednie wskaźniki finansowe, przy czym nie bierze się pod uwagę czynników jakościowych, szeroko wykorzystywanych w modelach scoringu

[41] A. Matuszczyk, *Dotychczasowe oraz nowe trendy w metodzie „credit scoring”,* Zeszyty Ekonomiczne Uniwersytetu Szczecińskiego, 2009, nr 548, s. 330.
[42] A. Matuszczyk, *Credit scoring,* op. cit., s. 156.
[43] G. Ignaciuk, op. cit., s. 23.
[44] T. Gestel, B. Baesens, op. cit., s. 104.
[45] A. Janc, M. Kraska, op. cit., s. 141.

kredytowego. Ponadto funkcje, w oparciu o które przeprowadza się analizy w modelach Z-score, w dużej części przypadków mają formę liniową, z kolei w credit scoringu, systemy mają postać zróżnicowaną[46].

Modele Z-score sprowadzają ocenę sytuacji finansowej przedsiębiorstwa zazwyczaj do analizy pojedynczego wskaźnika. Wskaźnik ten stanowi kombinację różnych wskaźników finansowych, połączonych w sposób ważony. Wartość funkcji Z wylicza się na podstawie informacji zawartych w sprawozdaniach finansowych. Technika ta pozwala na jednoznaczną ocenę kondycji przedsiębiorstwa[47].

1.3 System credit scoringu – etapy budowy

Podobnie, jak w przypadku większości systemów predykcyjnych, konstrukcja scoringu kredytowego opiera się na pewnym założeniu, że charakterystyki i ich zależności, świadczące o wypłacalności klientów, których sytuacja i ryzyko są znane bankowi z własnego portfela, stanowią równocześnie wyznacznik zdolności kredytowej statystycznie podobnych, przyszłych kredytobiorców. Zakłada się więc z dużym prawdopodobieństwem, że nowy klient zachowa się w podobny sposób jak znany bankowi kredytobiorca[48].

Procedura konstruowania systemu scoringowego składa się z kilku etapów. Poniżej został przedstawiony schemat budowy systemu scoringowego zaproponowany przez E.M Lewisa w 1992 roku. System ten składa się z czterech faz[49].

[46] A. Matuszczyk, *Credit scoring,* op. cit., s. 73.
[47] A. Tłuczak, *Zastosowanie dyskryminacyjnych modeli przewidywania bankructwa do oceny ryzyka upadłości* przedsiębiorstw, Zeszyty Naukowe Wyższej Szkoły Bankowej we Wrocławiu, 2013, nr 2(34), s. 425.
[48] A. Janc, M. Kraska, op. cit., s. 42.
[49] Ibidem, s. 42.

Tabela 1.1. Etapy budowy systemu scoringowego

I faza – koncepcyjna	II faza – projektowania	III faza – wdrożenia	IV faza – monitoringu
▪ Wybór grupy specjalistów ▪ Przyjęcie podstawowych celów	▪ Definicja dobrych i złych klientów ▪ Wybór populacji bazowej klientów ▪ Analiza danych i wybór właściwych charakterystyk ▪ Wybór typu metody estymacji modelu i przypisanie poszczególnym atrybutom właściwych wag, wyrażonych w punktach ▪ Konstrukcja tabeli scoringowej ▪ Ogólna statystyka i zatwierdzenie jakości, zdolności prognostycznych tabeli scoringowej	▪ Instalacja i instrukcja ▪ Ustalanie punktowej granicy oddzielającej grupę "dobrych" od "złych" ▪ Bezpieczeństwo ▪ Trening ▪ Zastosowanie systemu w praktyce	▪ Monitoring modelu

Źródło: A. Matuszczyk, *Ryzyko kredytowe*, w: Zarzadzanie ryzykiem w banku komercyjnym, pod red. M. Iwanicz-Drozdowksa, Wydawnictwo Poltext, Warszawa 2017, s. 148.

Faza I (koncepcyjna) polega na skompletowaniu zespołu specjalistów: analityków kredytowych, statystyków, informatyków, sprzedawców i ustaleniu celu modelu, czyli do czego ma być stosowany i jakich klientów ma badać[50].

Faza II (projektowania) należy do najbardziej złożonych. Definiuje się w niej klienta „złego" i „dobrego", przy czym definicja ta jest związana ściśle z wyznaczonym celem modelu. Na przykład, w przypadku modeli służących do określania prawdopodobieństwa braku terminowej spłaty zadłużenia, można przyjąć, że za złego klienta uważa się osobę, która nie wywiązała się z trzech kolejnych płatności, natomiast w przeciwnym wypadku uważany jest on za dobrego klienta[51]. Dodatkowo w tej fazie ustala się również tzw. populację bazową (*through the door population*), czyli przypadki, na podstawie których budowany będzie model. Różnie definiuje się jednak minimum liczby przypadków[52].

[50] A. Staniszewska, op. cit., s. 166.
[51] A. Matuszczyk, *Ryzyko kredytowe*, op. cit., s. 149.
[52] A. Staniszewska, op. cit., s. 166.

Kolejny etap budowy systemu scoringowego zakłada poddanie analizie wszystkich wyznaczonych przypadków w celu wyodrębnienia charakterystyk (tj. zmiennych cech dotyczących konkretnego podmiotu poddawanego scoringowi), zbieżnych do każdego klienta danej grupy, które określają przeciętnego, złego lub dobrego klienta banku. Etap ten przebiega na podstawie trzech procedur postępowania: kodowanie, wstępne zaliczenie oraz grupowanie[53].

Przez kodowanie rozumiemy identyfikacje wszystkich elementów wniosku kredytowego, które następnie mogą zostać wykorzystane w tworzeniu tabeli scoringowej. Wybrane elementy z wniosku kredytowego formułują listę tzw. charakterystyk, a więc zmiennych cech dotyczących określonego podmiotu branego do populacji bazowej. Po wydzieleniu charakterystyk, określa się tzw. atrybuty, czyli rzeczywiste informacje dotyczącej danej charakterystyki[54]. Następnie zdefiniowane charakterystyki i atrybuty zostają przetworzone do postaci umożliwiającej wygenerowanie pliku komputerowego.

Po wprowadzeniu do pliku i kodowaniu przechodzi się do procedury wstępnego zaliczania, która polega na określeniu liczby konkretnych zmiennych klasyfikowanych do grupy „dobrych" i „złych" klientów. Wstępne zaliczenie może posłużyć jako informacja, czy nie popełniono błędów w kodowaniu. Umożliwia to również sprawdzenie, czy liczby odzwierciedlają zgodną ze stanem faktycznym populację bazową. Jeżeli nie stwierdzono żadnych błędów, budowa tablicy skoringowej może być kontynuowana w kolejnym etapie, zwanym grupowaniem[55].

Tabela 1.2. Przykład kodowania

Wykształcenie	Podstawowe	Zawodowe	Średnie	Wyższe	Inne
Kod	0	1	2	3	4

Źródło: opracowanie własne.

Procedura grupowania ma rozwiązać dwa kluczowe problemy, które ujawniły się w procesie przeprowadzenia wstępnego zaliczenia[56]:

- liczba atrybutów – zbyt duża, aby mogła być zaimplementowana do systemu;
- liczba przypadków w poszczególnych atrybutach zbyt mała, aby można było wysnuć jakiekolwiek konkluzje.

[53] A. Matuszczyk, *Credit scoring,* op. cit., s. 92.
[54] Ibidem, s. 93.
[55] A. Janc, M. Kraska, op. cit., s. 39.
[56] A. Matuszczyk, *Ryzyko kredytowe*, op. cit., s. 151.

W grupowaniu stosowane się tabulogramy poprzeczne, o dużym stopniu szczegółowości. Zawarte są w nich takie informacje jak udział procentowy klientów zaliczonych do poszczególnych grup według określonego atrybutu i „szansa bycia dobrym", którą wylicza się w następujący sposób[57]:

$$Szansa\ bycia\ dobrym = \frac{\%\ klientów\ zaliczonych\ do\ grupy\ \text{dobrych}\ według\ danego\ atrybutu}{\%\ klientów\ zaliczonych\ do\ grupy\ \text{złych}\ według\ danego\ atrybutu}. \quad (1.1)$$

Znając udział procentowy oraz „szansę bycia dobrym" w ramach określonych atrybutów, dokonuje się grupowania atrybutów.

Następny etap w fazie projektowania przewiduje wybór metody estymacji modelu, za pomocą której zostaną obliczone właściwe wagi atrybutów. Tradycyjnie najczęściej stosowanymi technikami do budowy modelu są: analiza dyskryminacyjna, regresja liniowa, regresja logistyczna. Obecnie jednak zyskują na znaczeniu bardziej nowoczesne metody oparte między innymi o sieci neuronowe[58]. W tabeli 1.3 zostały zawarte popularne techniki statystyczne i niestatystyczne wykorzystywane do klasyfikacji kredytobiorców.

Tabela 1.3. Podział metod scoringowych

Metody stosowane w credit scoringu	
Statystyczne	**Niestatystyczne**
▪ Regresja liniowa	▪ Programowanie matematyczne
▪ Regresja logistyczna	(liniowe i całkowitoliczbowe)
▪ Analiza dyskryminacyjna	▪ Systemy eksperckie
▪ Drzewa klasyfikacyjne	▪ Sieci neuronowe
▪ Maszyny wektorów nośnych	▪ Algorytmy genetyczne

Źródło: Opracowanie własne.

Po ustaleniu wag przypisanych określonym atrybutom, następuje etap konstruowania końcowej tabeli scoringowej. Tabela taka zawiera wszystkie charakterystyki i powiązane z

[57] A. Janc, M. Kraska, op. cit., s. 94.
[58] A. Matuszczyk, *Credit scoring,* op. cit., s. 95.

nimi atrybuty oraz przypisaną im punktację[59]. W celu uzyskania kompletnej, przejrzystej tablicy skoringowej niezbędne staje się przetworzenie otrzymanych informacji w odpowiednią formę. W większości przypadków wagi obliczane są na podstawie poniższego wzoru[60]:

$$WoE = ln \frac{\begin{matrix}\%\ klientów\ zaliczonych\ do\ grupy\ \text{dobrych} \\ według\ danego\ atrybutu\end{matrix}}{\begin{matrix}\%\ klientów\ zaliczonych\ do\ grupy\ \text{złych} \\ według\ danego\ atrybutu\end{matrix}} \qquad (1.2)$$

Otrzymany współczynnik nazywany jest wagą dowodu (*weight of evidence*).

Po zatwierdzeniu jakości i skuteczności prognostycznej tabeli scoringowej następuje faza wdrożenia modelu. W fazie tej kluczowym elementem jest właściwy wybór tzw. punktu odcięcia (*cut-off-point*), rozdzielającego grupę klientów złych od dobrych. Pożyczkodawca musi określić, na jakim poziomie ustali punkt progowy, uwzględniając przy tym poziom ryzyka, które jest w stanie zaakceptować. Istnieje jednak problem, na jakim poziomie ustalić punkt odcięcia, gdyż ważna jest nie tylko minimalizacja strat, ale również utracone przychody banku[61]. Generalnie określenie wartości punktu odcięcia zależy w głównej mierze od prowadzonej polityki kredytowej oraz jakości aktywów banku.

Ostatnim etapem budowy systemu credit scoringu jest faza monitoringu (konieczna przy każdym rodzaju modelu). Model powinien podlegać ciągłemu monitoringowi po fazie wdrożenia. Analizuje się, czy wystąpiło jakieś zdarzenie losowe lub pojawiło się/eskalowało określone zjawisko, które może prowadzić do zmian w populacji kredytobiorców i tym samym odróżnia ją od rzeczywistej populacji bazowej (np. zmiany demograficzne). Model jest rewidowany również z powodu zmian w polityce kredytowej, to znaczy przy łagodzeniu lub zaostrzaniu oceny wiarygodności kredytowej klientów[62].

[59] Ibidem, s. 98.
[60] A. Janc, M. Kraska, op. cit., s. 54.
[61] A. Matuszczyk, *Ryzyko kredytowe*, op. cit., s. 154.
[62] A. Staniszewska, op. cit., s. 172.

Tabela 1.4. Końcowa tablica scoringowa

Status mieszkaniowy	Właściciel 45	Lokator 18	Inni 24	Brak informacji 30			
Posiadanie rachunków	Bieżący + terminowy 50	Terminowy 31	Bieżący 32	Bieżący + terminowy + inne 49	Bieżący lub terminowy + inne 23	Żaden nie podany 15	Brak informacji 15
Zawód	Wolny zawód 29	Menadżer 28	Urzędnik 25	Producent 15	Sprzedawca /kierowca 22	Inni 15	Brak informacji 24
Okres zatrudnienia u ostatniego pracodawcy	<1,5 roku 15	1,5-2,4 lat 22	2,5-4,4 lat 26	4,5-9,4 lat 26	9,5-12,4 lat 29	Brak informacji 23	
Ocena poprzednich rachunków	Niezadow. 0	Nowy 55	Brak ratingu 65	Zadow. 87	Brak informacji 47		
Posiadanie kart bankowych	Karta kredytowa 19	Inne karty 0	Brak kart 0	Brak informacji 8			
Najgorszy rating biura kredytowego	Brak informacji w b.k. 15	Negatywne -33	1 lub 2 zadow. 24	3 lub więcej zadow. 30	Brak informacji 0		
Towarzystwo ratalne	Tak 0	Nie 36	Bez kredytu 36	Brak informacji 20			

Źródło: A. Janc, M. Kraska, *Credit-scoring. Nowoczesna metoda oceny zdolności kredytowej*, Bibloteka Menadżera i Bankowca, Warszawa 2001, s. 55.

Rozdział 2. Statystyczne i eksploracyjne metody klasyfikacji danych

Modele statystyczne i eksploracji danych należą do najbardziej popularnych modeli oceny ryzyka kredytowego. Modele te pozwalają na zaklasyfikowanie kredytobiorcy do odpowiedniej grupy ryzyka lub określenia prawdopodobieństwa jego niewypłacalności przy użyciu obiektywnych i ilościowych zmiennych. W tym celu do badania wiarygodności kredytowej wykorzystuje się modele z jakościową zmienną zależną. Ryzyko kredytowe najczęściej wyraża się za pośrednictwem zmiennej binarnej lub zmiennych wielomianowych uporządkowanych[63].

2.1 Regresja logistyczna

Najbardziej popularną i ogólnie przyjętą w zagadnieniu klasyfikacji metodą modelowania zjawisk, gdzie prognozie podlega zmienna jakościowa, mogąca przyjmować wyłącznie dwie wartości, jest regresja logistyczna. Swoją popularność zawdzięcza ona przede wszystkim możliwości intuicyjnej interpretacji ocen parametrów regresji, które po nieskomplikowanym przekształceniu możemy rozumieć jako iloraz szans zajścia prognozowanego zjawiska. Oprócz możliwości opisu modelowanego zjawiska może być stosowana jako narzędzie predykcyjne, umożliwiające określenie ryzyka wystąpienia przewidywanego zjawiska u konkretnej osoby[64].

Warto wspomnieć, że pierwszy pełny model regresji logistycznej został opracowany dopiero w 1971 roku. Opisu tego modelu dokonał D. J. Finney w pracy pt.: *Probit analysis*[65]. Jednak pierwsze prace na temat zastosowań funkcji logistycznej pojawiły się już pod koniec XIX wieku w środowisku statystyków pracujących nad opisem właściwości demograficznych[66].

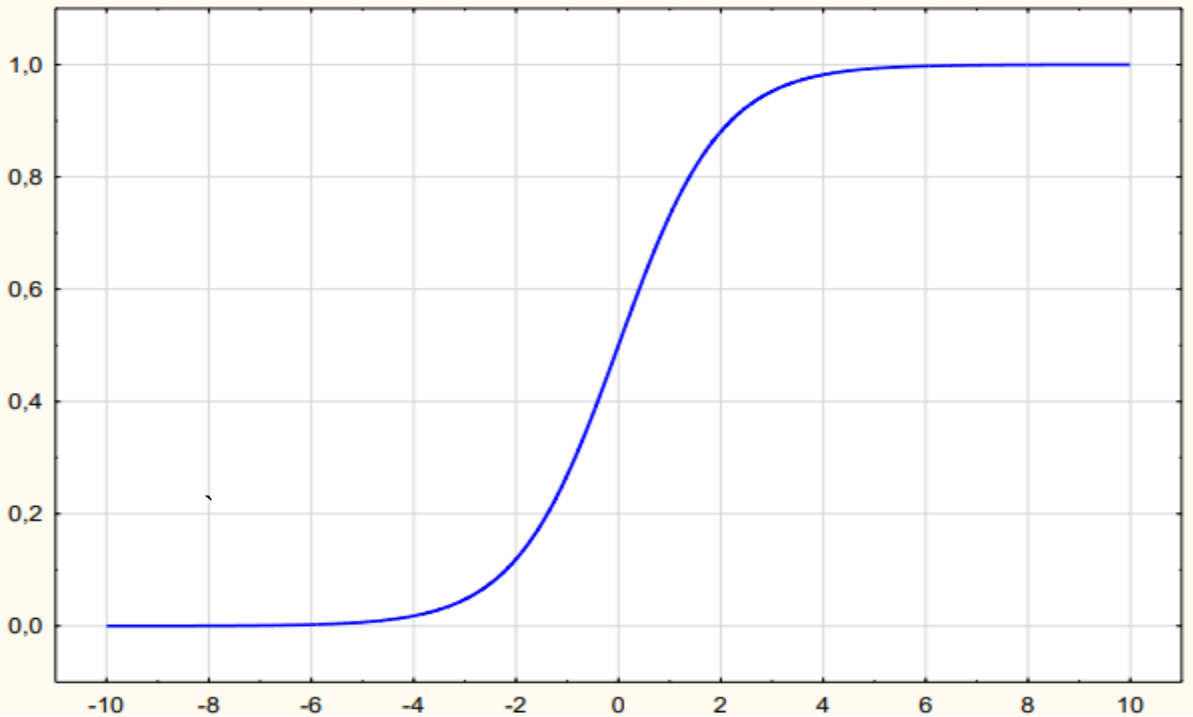
Model regresji logistycznej oparty jest o funkcję logistyczną postaci[67]:

$$f(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}. \qquad (2.1)$$

[63] M. Wójciak, op. cit., s. 69-70.
[64] T. Grochowiecki, *Zastosowanie regresji logistycznej do identyfikacji czynników ryzyka wystąpienia powikłań pooperacyjnych po jednoczesnej transplantacji trzustki i nerki,* StatSoft Polska, 2011, s. 30.
[65] *D.J. Finney, Probit Analysis. 3rd Edition, Cambridge University Press, Cambridge, 1971.*
[66] B. Danieluk, *Zastosowanie regresji logistycznej w badaniach eksperymentalnych*, Psychologai Społeczna, tom 5, 2010, nr 2-3 (14), s. 201.
[67] Dokumentacja algorytmów uczenia maszynowego, https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html (dostęp: 21.11.2019).

Funkcja logistyczna przyjmuje wartości od 0 do 1. Przebieg funkcji przedstawia wykres 2.1.

Wykres 2.1. Funkcja logistyczna



Źródło: opracowanie własne.

Własności funkcji logistycznej są następujące:

- Przyjmuje wartości od 0 do 1. Model może opisywać wartości prawdopodobieństwa, określającego ryzyko wystąpienia danego zjawiska lub szansę na niewystąpienie tego zjawiska (np. brak spłaty kredytu).
- Kształt funkcji przypomina rozciągniętą literę *S*. Oznacza to, że do osiągnięcia pewnej wartości progowej zmiany wartości funkcji są minimalne, potem gwałtownie wzrastają do 1 i utrzymują się na bardzo wysokim poziomie (bliskim 1).

Ogólnie rzecz ujmując, regresja logistyczna jest modelem matematycznym, umożliwiającym opisanie wpływu kilku zmiennych $X_1,X_2...,X_k$ na zmienną binarną Y.

Równanie regresji logistycznej umożliwia obliczenie warunkowego prawdopodobieństwa pojawienia się sukcesu w wyniku określonego czynnika i wyrażone jest następującym równaniem[68]:

$$P(Y = 1|X_1, X_2 \dots, X_k) = \frac{e^{\alpha+\sum_{i=1}^{k} \beta iXi}}{1 + e^{\alpha+\sum_{i=1}^{k} \beta iXi}}, \qquad (2.2)$$

gdzie:

$P(Y = 1|X_1, X_2 \dots, X_k)$ – warunkowe prawdopodobieństwo osiągnięcia przez zmienną objaśnianą wartości wyróżnionej (1) pod warunkiem uzyskania konkretnych wartości zmiennych objaśniających: $X_1, X_2 \dots, X_k$,

X_i – i-ta zmienna objaśniająca (i = 1, 2, …, k),

α – stała regresji logistycznej,

β_i – współczynnik regresji logistycznej dla i-zmiennej objaśniającej.

Dla obliczenia współczynnika P z równania (2.2), tj. wartości rozważanej funkcji, konieczne jest oszacowanie wielkości stałej regresji logistycznej (α) oraz współczynników regresji logistycznej (β_i). W przypadku modelowania za pomocą regresji liniowej stała regresji (α) oraz współczynnik regresji (β) szacowany jest metodą najmniejszych kwadratów. Metody tej nie można zastosować w modelowaniu z wykorzystaniem regresji logistycznej ze względu na brak liniowości rozkładu zmiennej objaśnianej. Współczynniki regresji logistycznej szacowane są metodą największej wiarygodności. Algorytm obliczeniowy metody największej wiarygodności polega na wielokrotnym estymowaniu każdego współczynnika regresji, tak by zmaksymalizować prawdopodobieństwo otrzymania takich wyników, jakie osiągnięto w analizowanej próbie[69].

Zatem nieznane parametry wektora β szacuje się na podstawie próby metodą największej wiarygodności. Jest to metoda iteracyjna.

Jeżeli wartości zmiennych objaśniających są podane, to rozkład zmiennej losowej $(Y_1, \dots, Y_n)$ zależy tylko od parametrów $\beta_1, \dots, \beta_k$. Ponieważ zmienne losowe $Y_1, \dots, Y_n$ są

[68] D. Kmieć, *Zastosowanie modelu logitowego do analizy czynników wpływających na bezrobocie wśród ludności wiejskiej*, Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego Ekonomika i Organizacja Gospodarki Żywnościowej, 2015, nr 110, s. 35.

[69] B. Danieluk, op. cit., s. 203.

niezależne, prawdopodobieństwo uzyskania zaobserwowanych wartości $y_1,\ldots,y_n$ w próbie wynosi[70]:

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) =$$

$$P(Y_1 = y_1)P(Y_1 = y_1) \cdot \ldots \cdot P(Y_1 = y_1) = \prod_{i=1}^{n} \mu_i{}^{y_i}(1-\mu_i)^{1-y_i}\,. \qquad (2.3)$$

Dla podanej próby powyższe prawdopodobieństwo jest funkcją parametrów $\beta_1,\ldots\ \beta_k$, określaną mianem funkcji wiarogodności próby:

$$L(\beta) = \prod\prod_{i=1}^{n} p_i^{y^i}(1-p_i)^{1-y_i}. \qquad (2.4)$$

Z uwagi na fakt, iż funkcja *L* osiąga maksimum w tych samych punktach co logarytm tej funkcji (tj. funkcja *lnL*), w praktyce określa się maksimum funkcji *lnL*. Maksimum to wyznacza się metodami rachunku różniczkowego, rozwiązując układ równań[71]:

$$\frac{\partial lnL}{\partial \beta_i} = 0, \qquad (2.5)$$

$$i = 0, \ldots, k.$$

W tym przypadku $lnL = \sum_{i=1}^{n}(y_i \ln(p_i) + (1-y_i)\ln(1-p_i)),$ (2.6)

$$\frac{\partial lnL}{\beta_i} = \sum_{i}^{n}(y_i - p_i)x_{ij.} \qquad (2.7)$$

Układ *k*+1 równań $\sum_{i=1}^{n}(y_i - p_i)x_{ij} = 0$ jest układem równań nieliniowych, który można rozwiązać, stosując iteracyjny algorytm Newtona-Raphsona.

Warto podkreślić, iż wartości oszacowań parametrów funkcji regresji logistycznej nie są interpretowalne. Dodatkowe wnioski odnośnie do modelowanego zjawiska, uzyskujemy wychodząc od modelu logitowego, który określany jest przez równanie:

$$logit(p) = log\frac{p}{(1-p)}\,. \qquad (2.8)$$

[70] D. Raczkiewicz, *Zastosowanie analizy regresji w reprezentacyjnych badaniach społeczno-gospodarczych*, „Econometrics", 2016, nr 1(51), s. 39.
[71] Ibidem s. 39.

Zaletą modelu logitowego jest zdolność do interpretacji parametrów e^{β_i}. W tym celu stosuje się pojęcie szansy (*odds*) określanej jako iloraz prawdopodobieństwa zajścia badanego zdarzenia oraz prawdopodobieństwa nie zajścia zdarzenia. W rozważanym modelu (2.2) szansę można przedstawić jako funkcję zmiennych objaśniających[72]:

$$\frac{p}{(1-p)} = \gamma(x_{1,}x_{2,}, \dots, x_k) = \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i). \qquad (2.9)$$

W przypadku wyrazu wolnego, wartość e^{β_0} jest interpretowana jako szansa zajścia zjawiska w grupie referencyjnej.

Wpływ przyrostu wartości zmiennych objaśniających o Δx_i (i=1,2,...,k) na szansę wystąpienia zjawiska można określić obliczając iloraz szans (*odds ratio*):

$$\psi(x_1, x_2, \dots, x_k; \Delta x_1, \Delta x_2 \dots, \Delta x_k) =$$

$$\frac{\gamma(x_1 + \Delta x_1, x_2 + \Delta x_2, x_k + \Delta x_k)}{\gamma(x_1, x_2, \dots, x_k)} = \exp\left(\sum_{i=1}^{k} \beta_i x_i\right). \qquad (2.10)$$

Jeżeli X_i (i = 1, 2, ...,k) jest zmienną zero-jedynkową, to e^{β_i} jest równy ilorazowi szans dla grupy, w której X_i = 1 lub grupy, w której X_i = 0, przy pozostałych zmiennych ustalonych. Natomiast, w przypadku gdy zmienna ta jest zmienną ilościową, to iloraz szans e^{β_i} mówi, jak zmieni się szansa, jeżeli zmienna X_i wzrośnie o 1 jednostkę przy pozostałych zmiennych niezmienioncyh[73].

Warto zauważyć, że w modelu regresji logistycznej nie wymaga się niektórych warunków koniecznych dla regresji liniowej. Wektor zmiennych niezależnych i reszty nie muszą mieć rozkładu normalnego oraz dopuszczalna jest heteroskedastyczność. Jednak konieczne jest spełnienie kilku innych założeń[74]:

- zmienna objaśniana musi być binarna, gdzie poziom określony jako ”1” reprezentuje pożądany wynik (sukces),

[72] B. Jackowska, *Efekty interakcji między zmiennymi objaśniającymi w modelu logitowym w analizie zróżnicowania ryzyka* zgonu, „Przegląd Statystyczny” R. LVII, 2011, Zeszyt 1-2, s. 25 -26.
[73] Ibidem, s. 26.
[74] J.Giemza, K. Zwierzchowska, *Wprowadzenie do modelu regresji logistycznej wraz z przykładem zastosowania w pakiecie statystycznym R do danych o pacjentach po przeszczepie nerk*i, praca licencjacka, Uniwersytet Warszawski Wydział Matematyki, Informatyki i Mechaniki, 2011, s. 17.

- zależność między logarytmem szans a wektorem zmiennych objaśniających powinna być liniowa,
- obserwacje muszą być niezależne – posługujemy się tym założeniem wyprowadzając postać funkcji wiarygodności,
- w danych nie może pojawiać się silna współliniowość – jest ona przyczyną problemów numerycznych,
- model musi być dobrze dopasowany, to znaczy składać się tylko z tych zmiennych niezależnych, które mają wpływ na zmienną zależną, oraz nie pomijać żadnej takiej zmiennej.

Ostatnie dwa założenia mają charakter raczej wskazówek niż warunków. Nie stosujemy ich do wyprowadzenia teorii regresji logistycznej, jednak model statystyczny, który ich nie spełnia, może prowadzić do nieprawidłowych wniosków.

2.2 Drzewa decyzyjne i rodziny klasyfikatorów

2.2.1 Drzewa decyzyjne

Drzewa decyzyjne pierwotnie, we wczesnych latach 60. XX w., były wykorzystywane w badaniach dotyczących psychologii i socjologii. W statystyce zagościły trochę później, bo po opublikowaniu prac Quinlana i Breimana[75].

Drzewa decyzyjne (*decision tree*) należą do metod statystycznych, w których dokonuje się klasyfikacji obserwacji próby statystycznej na grupy o podobnych właściwościach (stanowią odmianę hierarchicznej analizy skupień). W rezultacie otrzymuje się diagramy zwane drzewami klasyfikacji obserwacji statystycznych[76].

Z matematycznego punktu widzenia drzewo decyzyjne to acykliczny graf skierowany, w którym krawędzie nazywane są gałęziami, wierzchołki węzłami, wierzchołek nieposiadający rodzica korzeniem, a wierzchołki nieposiadające potomków liśćmi. Wszystkie węzły zawierają testy na atrybutach warunkowych zbudowane zgodnie z przyjętym kryterium podziału. Na podstawie wyników testów dokonywany jest podział danych w zależności od

[75] J. R. Quinlan, *Induction of decision trees*, „Machine Learning Journal", 1986; L. Breiman i in., *Classification and Regression Trees*, Wadsworth International Group, Monterey,Ca, 1984, vol. 1; cyt. za: M. Dramiński, *Algorytm indukcji reguł decyzyjnych w problemach klasyfikacji i wyboru cech w zadaniach wysokowymiarowych*, Rozprawa doktorska, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa 2007, s. 20.

[76] E. Matczak, W. Kozłowski, *Zastosowanie metody drzew klasyfikacyjnych w analizie aspiracji edukacyjnych rodziców*, Instytut Badań Edukacyjnych, Warszawa 2001, s. 6.

wartości ich cech (atrybutów), a każdy wynik testu reprezentowany jest przez gałęzie[77]. Aby zatem dokonać klasyfikacji nieznanego dla modelu przykładu przeprowadza się kolejne testy wartości, przesuwając się wzdłuż krawędzi po jednej ścieżce grafu, zaczynając od jego korzenia, a kończąc na właściwym liściu.

Rys. 2.1. Przykładowe drzewo klasyfikacyjne

Źródło: A. Matuszczyk, *Credit scoring*, Wydawnictwo CeDeWu, Warszawa 2018, s. 126.

Drzewo decyzyjne zazwyczaj konstruowane jest zstępująco - wychodząc od głównego węzła i dodając kolejne węzły potomne.

Podział w danym węźle przeprowadzany jest w oparciu o znajdujące się w nim elementy próby uczącej. Polega on na optymalnym podzieleniu próby na pewne części, które następnie przechodzą do węzłów dzieci. Próba ucząca znajdująca się w węźle charakteryzuje się określonym rozkładem klas, czyli tzw. różnorodnością klas reprezentowanych przez tę próbę. Racjonalny podział wymaga[78]:

1) zastosowania właściwej miary różnorodności klas w węźle,
2) zastosowania miary różnicy pomiędzy różnorodnością klas w danym węźle oraz klas w węzłach-dzieciach,
3) zastosowania algorytmu maksymalizacji tej różnicy.

[77] J. Kozak, P. Juszczyk, *Algorytmy do konstruowania drzew decyzyjnych w przewidywaniu skuteczności kampanii telemarketingowej banku,* Studia Informatica Pomerania, 2016, Nr 1, s. 51.
[78] M. Tyce, *Drzewa decyzyjne z użyciem pakietu R. Zastosowanie w badaniach występowania nawrotu choroby u pacjentek z nowotworem piersi,* praca licencjacka, Uniwersytet Warszawski Wydział Matematyki, Informatyki i Mechaniki, 2011, s. 8.

Rozważmy ustalony węzeł drzewa decyzyjnego (klasyfikującego) oznaczony jako *m*. Załóżmy przy tym, że próba losowa, której podpróba ucząca licząca n_m została zaklasyfikowana do węzła *m*. Ponadto liczba klas, w których skład mogą wchodzić obserwacje tej podpróby wynosi *g*. Wówczas prawdopodobieństwo, że dana obserwacja x_i zostanie przydzielona do klasy *k* w węźle *m* wyraża się wzorem[79]:

$$\hat{p}_{mk} = \frac{1}{n_m} \sum_{x_i \in R_m} I\,(y_i = k) = \frac{n_{mk}}{n_m}\ , \qquad (2.11)$$

gdzie R_m jest obszarem zawartym w przestrzeni obserwacji X $\supset R_m$ przez zestaw warunków, jakie musi spełnić obserwacja x_i, aby znaleźć się w węźle *m* oraz gdzie *I(A)* = 1, gdy warunek *A* jest spełniony i *I(A)* = 0, gdy warunek *A* nie jest spełniony.

W tym przypadku obserwacje znajdujące się w węźle *m* zostają zaklasyfikowane do takiej klasy *k*, dla której zachodzi:

$$k(m) = \arg(\max_k \hat{p}_{mk}), \qquad (2.12)$$

czyli takiej, która jest reprezentowana w tym węźle przez największą liczbę obserwacji z próby uczącej.

W drzewach decyzyjnych trzema najpowszechniej wykorzystywanymi miarami zanieczyszczenia (różnorodności) są: wskaźnik Giniego ($Q1_m(T)$), entropia ($Q2_m(T)$), oraz błąd klasyfikacji (($Q3_m(T)$). Miary te zostały przedstawione poniżej[80]:

$$Q1_m(T) = \sum_{k=1}^{g} \hat{p}_{mk}(1 - \hat{p}_{mk}) = 1 - \sum_{k=1}^{g} \hat{p}^2{}_{mk}\ , \qquad (2.13)$$

$$Q2_m(T) = -\sum_{k=1}^{g} \hat{p}_{mk} log \hat{p}_{mk}\ , \qquad (2.14)$$

$$Q3_m(T) = 1 - \hat{p}_{mk}(m). \qquad (2.15)$$

[79] M. Woldańska, *Zastosowanie drzew klasyfikacyjnych w badaniu zjawiska migracji klienta sieci telefonii komórkowej,* praca dyplomowa inżynierska, Politechnika Warszawska Wydział Elektroniki i Technik Informacyjnych Instytut Informatyki, 2013, s. 13.
[80] M. Tyce, *op. cit.*, s. 9.

Po określeniu miary różnorodności klas w węźle przy użyciu odpowiedniej metody, możemy przystąpić do wyznaczenia ułamka obserwacji, jakie przeszły z węzła m do jego prawego dziecka m_R i lewego dziecka m_L. Wynoszą one odpowiednio, dla dziecka prawego[81]:

$$\hat{p}_R = \frac{n_{m_R}}{n_m}, \qquad (2.16)$$

przy czym n_{m_R} stanowi liczbę obserwacji w węźle m_R , oraz analogicznie dla dziecka lewego:

$$\hat{p}_L = \frac{n_{m_L}}{n_m}. \qquad (2.17)$$

Należy przy tym zauważyć, że ponieważ rozważamy drzewa binarne, prawdopodobieństwo przejścia obserwacji do dziecka lewego wynosi:

$$\hat{p}_L = 1 - \hat{p}_R. \qquad (2.18)$$

Kolejną ważną wielkością w procesie budowy drzewa klasyfikacyjnego jest łączna miara różnorodności klas w węzłach – dzieciach rozważanego węzła m. Miara ta stanowi uśrednioną wartość różnorodności klas w dzieciach tego węzła, co przy zastosowaniu powyższych równości można przedstawić jako:

$$Q_{m_L,m_R} = \hat{p}_L Q_{m_L}(T) + \hat{p}_R Q_{m_R}(T). \qquad (2.19)$$

W oparciu o wszystkie powyższe zależności różnicę między różnorodnością klas w węźle – rodzicu i jego dzieciach formułuje się w sposób następujący:

$$\Delta Q_{m,m_L,m_R} = Q_m(T) - Q_{m_L,m_R}(T). \qquad (2.20)$$

W ostatnim etapie budowy drzewa klasyfikującego szukamy takiego podziału, który maksymalizowałby różnicę różnorodności klas między rodzicem a jego dziećmi. W praktyce sprowadza się to do rozpatrzenia wszystkich możliwych binarnych podziałów dla poszczególnych atrybutów i wybrania takiego, dla którego różnica między różnorodnością klas w węźle – rodzicu i różnorodnością klas w jego dzieciach będzie największa. Procedura

[81] J. Koronacki, J. Ćwik, *Statystyczne systemy uczące się,* Akademicka Oficyna Wydawnicza EXIT, Warszawa 2015, s. 137-138.

ta jest powtarzana dla każdego atrybutu danej próby, by ostatecznie wybrać optymalny podział dla najlepszego rozpatrzonego atrybutu[82].

W trakcie budowy drzewa decyzyjnego istnieje ryzyko nadmiernego dopasowania do zbioru trenującego. Drzewo decyzyjne może okazać się zbyt skomplikowane i odzwierciedlać przypadkowe zależności znajdujące się w zbiorze przykładów uczących, zwłaszcza, jeśli jest on "zaszumiony", dlatego oprócz samej konstrukcji drzewa stosuje się też metody służące do uproszczenia drzewa decyzyjnego w taki sposób, aby zachować zdolność modelu do uogólniania przykładów uczących i minimalizowania błędu rzeczywistego (powstającego podczas testowania obserwacjami nienależącymi do zbioru przykładów trenujących). Takie algorytmy potocznie nazywane są algorytmami przycinania[83].

Przez przycinanie rozumiemy usuwanie nieistotnych z punktu widzenia klasyfikatora składowych w pełni rozrośniętego drzewa. Proces ten polega na zastąpieniu jego wybranych poddrzew przez liście, którym nadaje się etykietę kategorii najczęściej występującej wśród powiązanych z nim przykładów uczących. Zmniejszenie rozmiaru drzewa, które powoduje uproszczenie jego konstrukcji, pomimo że skutkuje obniżeniem jakości klasyfikacji na zbiorze treningowym, podczas testowania daje znacznie lepsze wyniki. Drzewa decyzyjne uzyskane po przycięciu, poza zmniejszeniem błędu rzeczywistego, są łatwiejsze w interpretacji. Poza tym wymagają mniejszej wielkości pamięci i są skuteczniejsze podczas procesu klasyfikacji[84].

Implementację algorytmów przycinania drzew decyzyjnych można przeprowadzić na kilka sposobów[85]:

- przycinanie proste – dokonuje się poprzez zastąpienie przyciętego węzła węzłem zawierającym etykietę klasy-decyzji; jest to stosunkowo prosty mechanizm i dość popularny;
- przycinanie w trakcie wzrostu – przycinanie przeprowadza się podczas konstrukcji drzewa, z tym, że znalezienie kryterium stopu realizującego przycinanie w trakcie wzrostu z dobrymi rezultatami okazuje się trudne;

[82] M. Woldańska, op. cit., s. 15.
[83] Ł. Bujak, *Drzewa decyzyjne,* http://www.is.umk.pl/~duch/Wyklady/CIS/Prace%20zalicz/08-Bujak.pdf, s.12.
[84] G. Wilczewski, *InTrees: Modularne podejście do Drzew Decyzyjnych*, praca magisterska, Uniwersytet Mikołaja Kopernika Wydział Matematyki i Informatyki, Toruń 2008, s. 16-17.
[85] D. Przywara, *Drzewa decyzyjne, metody budowania, zastosowania*, Praca zaliczeniowa do kursu „Informatyka systemów autonomicznych", Wydział Elektroniki Politechniki Wrocławskiej, Wrocław 2007, s. 4.

- przycinanie od środka – jest to zastępowanie przyciętego węzła jednym z jego węzłów potomnych, ale nie jest usuwane całe drzewo, a jedynie jego korzeń.

Oprócz odpowiedniego algorytmu modyfikacji drzewa, należy jeszcze określić, w jaki sposób szacowany będzie współczynnik błędu rzeczywistego drzewa decyzyjnego, na podstawie którego podejmuje się decyzję o przycięciu poddrzewa. Wyróżnia się dwa główne podejścia[86]:

1. Przycinanie na podstawie przykładów trenujących – do obliczania współczynnika błędu rzeczywistego wykorzystywany jest ten sam zbiór przykładów. Daje to gorsze rezultaty, aczkolwiek przy małym zestawie przykładów takie rozwiązanie jest konieczne.
2. Przycinanie z oddzielnym zbiorem przycinania – do obliczenia współczynnika błędu rzeczywistego używany jest oddzielny zbiór przykładów niebiorący udziału w budowie drzewa. Zwykle przy dużej ilości danych używa się 2/3 przykładów do algorytmu rekurencyjnego konstruowania drzewa, natomiast pozostała 1/3 służy do określania błędów rzeczywistych w poszczególnych węzłach gotowego już drzewa decyzyjnego.

Najbardziej zaawansowanym algorytmem indukcji drzewa decyzyjnego jest CART (*Classification And Regression Tree*) zaproponowany przez Breimana i in. w 1984 roku[87]. Jest to nieparametryczny algorytm wykorzystujący standardowy schemat zstępującej metody konstrukcji drzewa. Cechą charakterystyczną algorytmu CART jest całkowite przeszukiwanie dziedziny w ramach wszystkich możliwych podziałów na dwa oddzielne i uzupełniające się podzbiory. Każdy z tych podziałów jest oceniany według określonego kryterium i stosowany jest najlepszy z nich. W metodzie CART, jako kryterium oceniające jakość podziałów przeważnie stosuje się miarę entropii i wskaźnik Giniego[88]. Algorytm CART wyróżnia się spośród innych algorytmów drzew decyzyjnych nadmiernym rozrostem i w efekcie koniecznością przycinania (*pruning*) poszczególnych gałęzi w celu zmniejszenia opisu liści. Cecha ta umożliwia porównanie modelu bardziej złożonego i modelu ze zredukowaną liczbą węzłów, bowiem czasami o jakości drzewa nie decyduje trafność predykcji, ale użyteczność

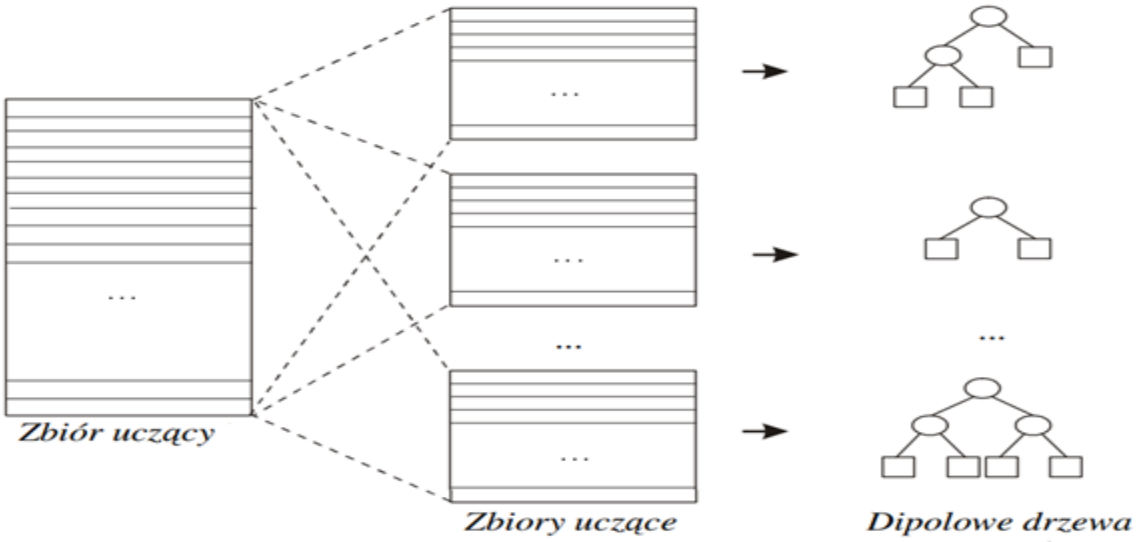
[86] Ibidem, s. 5.
[87] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Monterey,Ca, 1984.
[88] A. Hołda, *Wykorzystywanie drzew decyzyjnych w prognozowaniu upadłości przedsiębiorstw w branży budowlanej*, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, 2009, nr 796, s. 170.

wygenerowanych reguł. Dokonuje się sprawdzenia, jaka jest różnica między błędem klasyfikacji całego drzewa, a błędem klasyfikacji drzewa z usuniętą gałęzią[89].

2.2.2 Lasy losowe

Las losowy (*random forest*) to zbiór wielu klasyfikatorów (*ensemble classifier*), w ramach którego każdy pojedynczy klasyfikator jest drzewem decyzyjnym uczonym bez zatrzymywania (brak kryterium stopu oznacza, że w liściu znajdują się tylko próbki z tej samej klasy). Każdy klasyfikator będący częścią lasu losowego jest uczony na specjalnie wylosowanym dla niego zestawie danych *D*`, który powstaje przez wylosowanie próbek *n* razy ze zwracaniem ze wszystkich *N* próbek uczących[90].

Rys. 2.2. Konstrukcja lasu losowego



Źródło: M. Krętowska, *Lasy losowe – ocena jakości prognostycznej cech*, Zeszyty Naukowe Politechniki Białostockiej, Nr 2, 2007, s. 71.

Algorytm *Random Forest* zaproponowany przez Breimana w 2001 roku w swej podstawowej wersji składa się z trzech kroków[91]:

1. Wyznacz liczbę modeli bazowych *V* oraz liczbę zmiennych *K*.
2. Wykonaj dla każdego $v = 1, ..., V$ poniższe kroki:
   - ✓ Wylosuj próbę uczącą U_V ze zbioru uczącego *U*.

[89] M. Łapczyński, *Drzewa klasyfikacyjne w badaniach satysfakcji i lojalności klientów*, Statsoft Polska, 2003, s. 95.
[90] P. Płoński, *Zastosowanie wybranych metod przekształcania i selekcji danych oraz konstrukcji cech w zadaniach klasyfikacji i klasteryzacji,* Rozprawa doktorska, Politechnika Warszawska, Warszawa 2016, s. 18.
[91] L. Breiman, *Random Forests*, „Machine Learning", 2001, nr 45, s. 5-32.

- ✓ Skonstruuj maksymalne drzewo D_V w oparciu o próbę U_V, losując w każdym węźle drzewa K zmiennych, wśród których najlepsza dobierana jest do modelu.

3. Dokonaj prognozy modelu zagregowanego dla przykładów x_i za pomocą modeli bazowych $D_1,\ldots, D_V$, posługując się następującą metodą głosowania[92]:

$$\widehat{D}^* = arg \max_s \sum_{v=1}^{V} I(\widehat{D}_v(x_i) = P_s). \qquad (2.21)$$

Model zagregowany $\widehat{D}^*$ przydziela obserwacje x_i do tej klasy P_s (gdzie $s = 1,\ldots,u$), którą wskazała największa liczba modeli bazowych $D_1,\ldots, D_V$.

Algorytm lasów losowych pozwala na uniknięcie problemów związanych z modelowaniem pojedynczym drzewem klasyfikacyjnym, takich jak niestabilność, czyli duża wrażliwość na małe zmiany w próbie, a także zmniejsza problem „przeuczenia" i potrzebę „przycinania" drzewa[93].

W lasach losowych poszczególne klasyfikatory trenowane są na próbach bootstrapowych (losowanych ze zwracaniem), z których każda składa się średnio z około $^2/_3$ spośród obserwacji pierwotnej próby. Obserwacje spoza pseudopróby uczącej (tak zwane przykłady *out-of-bag*, OOB) można więc zastosować do szacowania różnych wielkości przydatnych w trakcie konstruowania komitetu (tj. zbioru klasyfikatorów), jak i na etapie analizy danych i predykcji. W oparciu o przykłady OOB można wyliczyć m. in. oszacowanie błędu klasyfikacji poszczególnych predykatorów, błędu klasyfikacji komitetu (w takim przypadku każda z obserwacji pierwotnej próby jest klasyfikowana tylko przez te klasyfikatory, w których trenowaniu nie uczestniczyła), miarę istotności zmiennych wchodzących w skład modelu oraz bliskości obserwacji, którą można wykorzystać do identyfikowania obserwacji odstających (ostatnia miara stosowana jest tylko w przypadku użycia drzew jako klasyfikatorów)[94].

[92] M. Walesiak, E. Gatnar, *Statystyczna analiza danych z wykorzystaniem programu R,* Wydawnictwo Naukowe PWN, Warszawa 2012, s. 267-268.
[93] R. Górska, P. Staszkiewicz*, Zastosowanie algorytmu lasów losowych do prognozowania modyfikacji opinii biegłego rewidenta*, „Zarządzanie i Finanse - Journal of Management and Finance", 2017, vol. 15, nr 3, s. 342.
[94] Z. Branicka, *Metody konstrukcji oraz symulacyjne badanie właściwości jednorodnych i niejednorodny1ch komitetów klasyfikatorów*, Praca magisterska, Uniwersytet Warszawski Wydział Matematyki, Informatyki i Mechaniki, 2001, s. 10.

2.2.3 Drzewa wzmacniane gradientowo

Wzmacnianie (*boosting*) dotyczy każdej metody zespołowej łączącej kilka słabych klasyfikatorów w jeden silny klasyfikator. Podstawową koncepcją w większości technik wzmacniania jest sekwencyjne uczenie predyktorów w taki sposób, że każdy kolejny próbuje korygować poprzednika[95]. Istnieje wiele metod wzmacniania, do jednych z najbardziej popularnych algorytmów tego typu zalicza się drzewa z wzmacnianiem gradientowym (*gradient boosted trees*) oraz algorytm XGBoost (*extreme gradient boosting*).

Metoda wzmacniania, zgodnie z propozycją Freunda i Schapire z 1996 roku, powoduje utworzenie modelu zagregowanego, który jest oparty na strukturze szeregowej[96]. Inaczej mówiąc, zawartość próby uczącej U_v zależy od wyników predykcji modelu znajdującego się poprzednio w szeregu, tj. D_{v-1}. To „wzmocnienie" uzyskuje się dzięki posłużeniu się podwójnym systemem wag. Pierwszy system polega na tym, że obserwacje ze zbioru U, które zostały niewłaściwie sklasyfikowane przez model D_v, otrzymują wyższe wagi. Natomiast drugi system dotyczy modeli bazowych i polega na przypisaniu każdemu z modeli wagi adekwatnie do jego błędu predykcji $e(D_v)$, tj.:

$$ln\left(\frac{1-e(D_v)}{e(D_v)}\right). \qquad (2.22)$$

Oznacza to, że mniejsze wagi są przydzielane modelom mniej dokładnym, a większe wagi tym modelom, które bardziej precyzyjnie klasyfikują obserwacje ze zbioru U[97].

W przypadku algorytmu drzew wzmacnianych gradientowo nie aktualizujemy wag przykładów po każdym przebiegu, lecz próbujemy dopasować predyktor do błędu resztowego popełnionego przez poprzedni predyktor[98]. Minimalizacja funkcji błędu dokonywana jest metodami gradientowymi.

[95] A. Geron, *Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow*, Wydawnictwo Helion, Gliwice 2018, s. 197.
[96] Y. Freund, R.E Schapire, *A decision-theoretic generalization of on-line learning ana an application to boosting*, "Journal of Computer and System Sciences", 1996, nr 55, s. 119-139.
[97] M. Walesiak, E. Gatnar, op. cit., s. 264-265.
[98] A. Geron, op. cit., s. 200.

Biorąc pod uwagę dane treningowe $X = \{x_i\}_{i=1}^{N}$ z $x_i \in R^D$ oraz ich etykiety $Y = \{y_i\}_{i=1}^{N}$ z $y_i \in \{0,1\}$, zadaniem algorytmu jest wybranie funkcji klasyfikacji *F(x)* minimalizującej agregację określonych funkcji straty $\mathcal{L}(y_i, F(x_i))$[99]:

$$F^* = arg \min_F \sum_{i=1}^{N} \mathcal{L}(y_i, F(x_i)) \, . \qquad (2.23)$$

Algorytm wzmacniania gradientowego rozważa szacowanie funkcji *F* w addytywnej formie:

$$F(x) = \sum_{m=1}^{T} f_m(x) \, , \qquad (2.24)$$

gdzie:

T – jest liczbą iteracji,

$f_m(x)$ – zaprojektowane są w sposób przyrostowy,

m – etap nowo dodawanej funkcji,

f_m – wybierana jest w celu zoptymalizowania zagregowanej straty przy zachowaniu $\{f_j\}_{j=1}^{m-1}$ stałej.

Każda funkcja f_m należy do zestawu sparametryzowanych bazowych klasyfikatorów. Niech θ oznacza wektor parametrów bazowych klasyfikatorów (w algorytmie drzew wzmacnianych gradientowo bazowymi klasyfikatorami są drzewa decyzyjne). Na wybór wektora θ składają się parametry reprezentujące strukturę drzewa, takie jak funkcja podziału w każdym węźle wewnętrznym, próg podziału każdego węzła itp.

Na etapie *m* formułujemy przybliżoną funkcję straty:

$$\mathcal{L}(y_i, F_{m-1}(x_i) + F_m(x_i)) \approx \mathcal{L}\big(y_i, F_{m-1}(x_i)\big) + g_i f_m(x_i) + \frac{1}{2} f_m(x_i)^2 \, , \qquad (2.25)$$

gdzie:

$F_{m-1}(x_i) = \sum_{j=1}^{m-1} f_j(x_i)$

i $g_i = \frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)} | F(x_i) = F_{m-1}(x_i).$

[99] S. Si, H. Zhang, S. S. Keerthi, D. Mahajan, I. S. Dhillon, C. J. Hsieh, *Gradient Boosted Decision Trees for High Dimensional Sparse Output*. In ICML, 2017, s. 3184.

Należy wyznaczyć F_m, aby zminimalizować prawą stronę równania (2.25), co można zapisać jako następujący problem minimalizacji:

$$arg \min_{f_m} \sum_{i=1}^{N} \frac{1}{2} (f_m(x_i) - g_i)^2 \ . \qquad (2.26)$$

Warto podkreślić, że w tym przypadku tylko kierunek optymalizacji gradientem jest dopasowany, a odpowiedni współczynnik udziału każdego drzewa w zespole (parametr kurczliwości) jest zwykle kształtowany w f_m przed dodaniem go do F_{m-1}. Zaletą takiego podejścia wzmacniania gradientem jest to, że zmienia się tylko wyrażenie gradientu dla określonych funkcji strat, podczas gdy reszta procedury, a zwłaszcza etap indukcji drzewa decyzyjnego, pozostaje taki sam dla tych funkcji strat.

2.2.4 Algorytm XGBoost

Algorytm *Extreme Gradient Boosting* (*XGBoost*), będący pochodną drzew decyzyjnych i lasów losowych, został opracowany w 2014 r., upowszechniony zaś w 2016 roku przez Tianqi Chen i Carlosa Guestrin[100]. XGBoost jest uogólnioną implementacją wzmacniania gradientu, która obejmuje termin regularyzacji. Używany jest do zwalczania nadmiernego dopasowania, a także jest wsparciem dla arbitralnych różnicowych funkcji straty.

Do najważniejszych cech algorytmu XGBoost zaliczamy[101]:

1. Model XGB nie ma postaci analitycznej – jest formułowany poprzez numeryczną optymalizację funkcji błędu resztowego w kolejnych iteracjach.

2. Algorytm ma postać modelu złożonego (*ensemble*) – jest budowany w oparciu o regresyjne lub klasyfikacyjne drzewa decyzyjne typu CART partycjonujące przestrzeń przykładów metodą dziel-i-rządź, z zamiarem uzyskania maksymalnego porządku danych (według zadanych kryteriów). XGBoost uczy szereg drzew, z których każde kolejne drzewo dopasowuje się do błędu resztowego popełnionego przez poprzedni predyktor, osiągając lepszą skuteczność.

[100] T. Chen, C. Guestrin, *XGBoost: A Scalable Tree Boosting System* [w:] Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, ACM Press, New York, 2016, s. 785-794.
[101] F. Wójcik, *Prognozowanie dziennych obrotów przedsiębiorstwa za pomocą algorytmu XGBoost – studium przypadków,* Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, 2018, nr 375, s. 128-129.

3. Każde kolejne drzewo poddawane jest regularyzacji, (jest to swego rodzaju kara nakładana na model za zbyt dużą liczbę ostatecznych segmentów obserwacji, czyli liści w drzewie decyzyjnym). W celu ograniczenia przetrenowania modelu (regularyzacji) stosuje się zwykle metodę regresji grzbietowej lub regresję metodą LASSO.

4. Obserwacje treningowe w modelu dobierane są na podstawie losowania ze zwracaniem (tzw. bootstrap) z pierwotnego zbioru *D*, z założeniem rozkładu jednostajnego – tj. każda próbka ma takie same szanse znaleźć się w podzbiorze. Przykłady, które nie zostały przydzielone do zbioru treningowego, służą jako zbiór testowy w danej iteracji (tzw. *out of bag error*) i wartości resztowe powstałe w ten sposób są estymatorami błędu rzeczywistego.

5. Wartość błędu resztowego popełnionego przez klasyfikator (drzewo decyzyjne) dla każdego elementu badanej populacji jest rejestrowana i wykorzystywana przez następne drzewo, by poprawić dotychczasowe wyniki. Funkcja błędu drzewa numer *t* w stosunku do wartości oczekiwanych zmiennej objaśnianej wygląda następująco[102]:

$$\mathcal{L}^t \approx \sum_{i=1}^{N} l\left(y_i, \hat{y}_l{}^{t-1} + f(x_i)^t + \Omega(f^t)\right), \qquad (2.27)$$

gdzie:

t – numer iteracji (numer kolejny drzewa)

l (x,y) – funkcja penalizująca błąd, np. RMSPE,

y_i – wartość zmiennej objaśnianej w *i*-tej obserwacji treningowej,

$\hat{y}_l{}^{t-1}$ – prognozowana przez drzewo numer *t*–1 (poprzednie w kolejności) wartość *i*-tej zmiennej objaśnianej,

$f(x_i)^t$ – prognoza wartości *i*-tej zmiennej endogenicznej, uzyskana przez drzewo *t*,

$\Omega(f^t)$ – funkcja regularyzacji drzewa numer *t*.

Z powyższego wynika, iż algorytm XGBoost jest nie tylko algorytmem złożonym, ale także iteracyjną procedurą typu boosting (stąd nazwa), której głównym zadaniem jest korekcja predykcji przy wykorzystaniu błędów poprzednich iteracji.

[102] T. Chen, C. Guestrin, *XGBoost: A Scalable Tree Boosting System,*[w:] Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, ACM Press, New York, 2016, s. 786.

2.3 Maszyny wektorów wspierających

Początki metody maszyn wektorów wspierających (*suport vector machine* – SVM) sięgają lat 70. XX wieku i wczesnych prac Vapnika i Chervonenkisa[103]. Jednak inspiracji dla jej powstania można doszukiwać się już w latach 50. Wówczas pojawiła się również idea rozdzielania klas przy użyciu liniowej granicy decyzyjnej (hiperpłaszczyzny), której położenie byłoby optymalne względem do zaobserwowanej próby uczącej. W 1958 roku Rosenblatt[104] zaproponował algorytm poszukujący hiperpłaszczyzny, przy której sumaryczna odległość niewłaściwie sklasyfikowanych obserwacji od granicy decyzyjnej byłaby możliwie najmniejsza[105].

Dzięki metodzie SVM możliwe jest przyporządkowanie wejściowego zbioru danych do wcześniej definiowanych klas. Algorytm ten do separacji klas wykorzystuje hiperpłaszczyznę, która je rozdziela wraz z maksymalnym możliwym marginesem. Obserwacje znajdujące się najbliżej hiperpłaszczyzny nazywane są „wektorami wspierającymi" (*support vectors*) i są najbardziej istotnymi składowymi klasyfikacji. W ramach przeprowadzania tej metody określana jest przynależność każdego punktu do danej klasy. Wynik klasyfikacji zależy od szerokości granicy rozdzielającej poszczególne klasy, czyli przywołanego powyżej marginesu. Im większy jest margines, tym mniejszy błąd klasyfikacji[106]. Metoda wektorów wspierających jest użyteczna zarówno dla rozwiązań liniowych jak i nieliniowych. Co więcej, metoda ta ma również swoje zastosowanie w problemach regresji[107].

Celem klasyfikatora SVM jest wyznaczenie optymalnej hiperpłaszczyzny rozdzielającej dane separowalne liniowo w przestrzeni $\mathbb{R}^n$, a także określenie dwóch równoległych i znajdujących się w tej samej odległości hiperpłaszczyzn tworzących margines klasyfikacji[108].

---

[103] V.V. Vapnik, A.Y. Chervonenkis, *Theory of pattern recognition*, "Nauka", 1974, nr 107.
[104] F. Rosenblatt , *The perceptron: a probabilistic model for information*, "Psychological Review", 1958.
[105] D. Gąska, *Zastosowanie metody SVM do oceny ryzyka bankructwa i prognozowania upadłości przedsiębiorstw*, „Śląski Przegląd Statystyczny", 2013, nr 11(17), s. 292-293.
[106] A. Marcinkowska-Ochtyra, *Ocena przydatności obrazów hiperspektralnych APEX oraz maszyn wektorów nośnych (SVM) do klasyfikacji roślinności subalpejskiej i alpejskiej Karkonoszy,* Rozprawa doktorska, Uniwersytet Warszawski, 2016, s. 37.
[107] A. Jakubczyk-Gałczyńska, A. Kristowski, R. Jankowski, *Idea zastosowania sztucznej inteligencji w prognozowaniu wpływu drgań komunikacyjnych na odpowiedź dynamiczną budynków mieszkalnych*, XI Konferencja Nowe Kierunki Rozwoju Mechaniki, Sarbinowo Morskie 2015, s. 6.
[108] Ibidem, s. 38.

Niech x_i będzie wektorem wejściowym zaś y_i oznaczeniem klasy przyjmującym wartości {-1,1}. Analizowany jest zbiór uczący jako para (x_i, y_i) dla i=1, 2…, p, $x_i \in R^n$. Przyjmijmy założenie, że klasy y_i są liniowo separowane. Wówczas funkcja *g(x)* określona równaniem (2.28) będzie hiperpłaszczyzną rozdzielającą obie klasy[109]:

$$g(x) = w^T x + b = 0\,, \qquad (2.28)$$

gdzie: $w = [w_1, w_2, w_3, \dots, w_N]^T$, $x = [x_1, x_2, x_3, \dots, x_N]^T$.

Jeżeli spełnione są warunki (2.29), to optymalną hiperpłaszczyznę, która maksymalizuje margines separacji możemy wyznaczyć równaniem (2.30), zaś odległość *odl* wybranego obiektu *x* od optymalnej hiperpłaszczyzny równaniem (2.31).

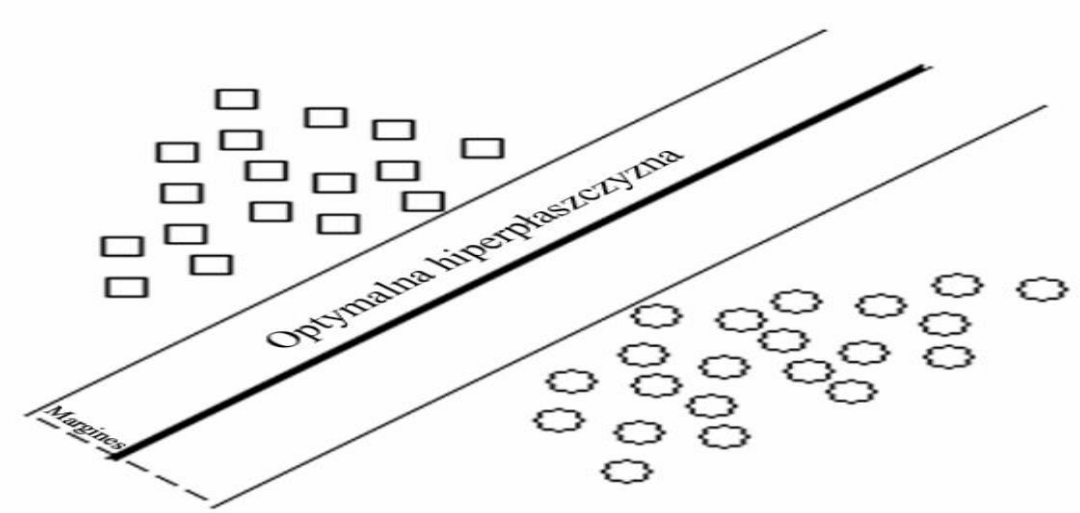
$$\begin{cases} w^T x + b > 0 & dla \quad y_i = 1 \\ w^T x + b < 0 & dla \quad y_i = -1 \end{cases}, \qquad (2.29)$$

$$g(x) = w_0^T x + b_0 = 0\,, \qquad (2.30)$$

$$odl(x) = \frac{g(x)}{\|w_0\|}\,. \qquad (2.31)$$

Ilustrację graficzną tworzonych hiperpłaszczyzn metodą wektorów wspierających pokazano na rys. 2.3.

Rys. 2.3. Wizualizacja maszyny wektorów wspierających



Źródło: K. Gałda, *Zastosowanie algorytmów genetycznych do optymalizacji modelu SVM procesu stalowniczego* , Praca magisterska, Politechnika Śląska, Katowice 2009, s. 20-21.

[109] A. Duraj, *Wykrywanie wyjątków przy użyciu wektorów nośnych,* Zeszyty Naukowe WSInf, 2017, vol. 16, nr 1, s. 58.

Ustanowienie możliwie największego marginesu między dwiema klasami wymaga maksymalizacji odległości punktów (najbliżej położonych) od hiperpłaszczyzny. Taki cel może być określony w następujący sposób:

$$\max_{w,b} \quad \min \ \{ \|x - x_i\| \ : \ w^T x + b = 0, \ \ i = 1, \dots, N\}, \qquad (2.32)$$

przy czym $w^T x + b = 0$, to równanie hiperpłaszczyzny, gdzie przez *w* oznaczamy wektor prostopadły do niej, a przez *b* odległość od środka układu współrzędnych. Przeprowadzając kilka przekształceń otrzymujemy funkcję celu, określoną jako[110]:

$$\min_{w,b} \quad \tau(w) \frac{1}{2} \|w\|^2 \ . \qquad (2.33)$$

Zakładając, że:

$$y_i[w^T x_i + b] \geq 1, \qquad i = 1, \dots, N, \qquad (2.34)$$

a następnie wykorzystując metodę mnożników Lagrange'a, uzyskujemy Lagrangiana zdefiniowanego jako:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{N} \alpha_i (y_i [x_i{}^T w + b] - 1) \ , \qquad (2.35)$$

gdzie $\alpha_i > 0$ są mnożnikami Lagrange'a. W takim przypadku naszym celem jest maksymalizacja Lagrangiana *L* ze względu na α_i i minimalizacje względem *w* i *b*. To prowadzi do warunków, w których pochodne *L* względem tych współczynników zanikają i otrzymujemy:

$$w = \sum_{i=1}^{N} \alpha_i y_i \, x_i \ , \qquad (2.36)$$

$$\sum_{i=1}^{N} \alpha_i \, y_i = 0 \ . \qquad (2.37)$$

Wektory x_i, dla których $\alpha_i > 0$, nazywane są wektorami wspierającymi bądź podpierającymi.

Należy jednak odnotować, że rozwiązanie powyższego zadania może nie być możliwe z uwagi na brak liniowej separowalności. Wówczas równanie (2.35) dla danych

[110] K. Gałda, *Zastosowanie algorytmów genetycznych do optymalizacji modelu SVM procesu stalowniczego*, Praca magisterska, Politechnika Śląska, Katowice 2009, s. 20-21.

nieseparowalnych liniowo można zapisać w formie równania (2.38), gdzie ξ_i oznacza nieujemną zmienną dopełniającą, zaś φ stanowi wagę wybraną przez użytkownika, określającą postępowanie wobec błędów testowania w odniesieniu do wyznaczonego marginesu[111]:

$$min\left(\frac{1}{2}w^T w + \varphi \sum_{i=1}^{N} \xi_i\right). \qquad (2.38)$$

W celu rozwiązania powyższego problemu optymalizacji można zastosować metodę maksymalizacji dualnej funkcji Lagrange'a $-L_D$, opisanej odpowiednio przez równanie (2.39), tj.:

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j\,, \qquad (2.39)$$

z ograniczeniami:

$$0 \leq \alpha_i \leq \varphi \qquad i = 1, \ldots, N,$$

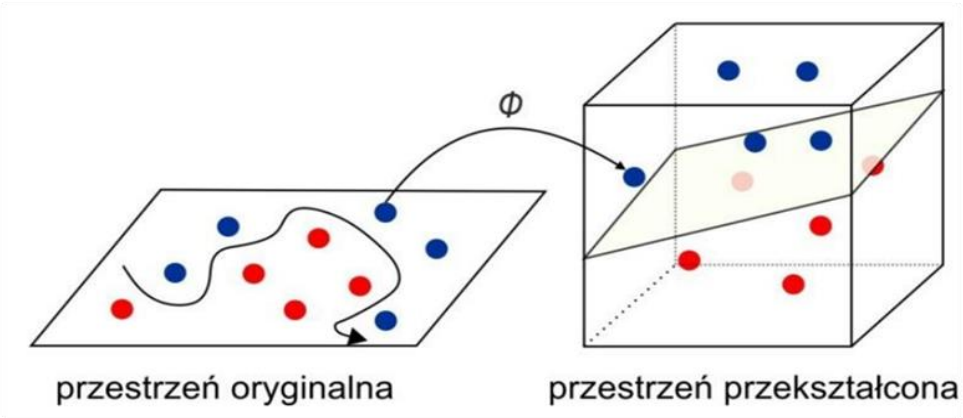
$$\sum_{i=1}^{N} \alpha_i y_i = 0. \qquad (2.40)$$

Należy zauważyć, że postać funkcji jest identyczna jak w przypadku funkcji problemu separowanych liniowo, jednak ograniczenie jest trochę inne.

W najbardziej zaawansowanej wersji metoda maszyn wektorów wspierających łączy ideę klasyfikacji liniowej z technikami jądrowymi. Chociaż przedstawiona uprzednio wersja metody pozwala na dyskryminację w przypadku, gdy klasy nie są liniowo separowalne, jeszcze lepszym wariantem okazał się koncept, aby przestrzeń obserwacji X przekształcić nieliniowo do określonej przestrzeni o większym (potencjalnie nieskończonym) wymiarze $\varphi(\mathrm{X})$, określonej przez operator φ, a następnie zastosować liniowy model dyskryminacji w tej bogatszej przestrzeni[112]. Na rysunku 2.4 zamieszczona jest ilustracja poglądowa do tego pomysłu.

[111] A. Duraj, op. cit., s. 59.
[112] D. Gąska, op. cit., s. 296-297.

Rys. 2.4. Transformacja danych wejściowych do nowej przestrzeni wielowymiarowej



Źródło: A. Marcinkowska-Ochtyra, *Ocena przydatności obrazów hiperspektralnych APEX oraz maszyn wektorów nośnych (SVM) do klasyfikacji roślinności subalpejskiej i alpejskiej Karkonoszy,* Rozprawa doktorska, Uniwersytet Warszawski, 2016, s.39.

Ideą metody SVM wykorzystującej techniki jądrowe jest konstrukcja optymalnej hiperpłaszczyzny w pewnej wielowymiarowej przestrzeni cech *Z*, która jest nieliniowym produktem poszczególnych funkcji jądrowych φ(.) wybranych a priori. Metoda SVM znajduje liniową hiperpowierzchnię w wysokowymiarowej przestrzeni cech *Z*, wyznaczając przy tym możliwie największy margines rozdzielający. W metodzie SVM optymalizacja marginesu dokonywana jest przez znalezienie możliwie najlepszego parametru kary *C* oraz parametrów funkcji jądrowych[113].

Aby uniknąć pracy na wielowymiarowej przestrzeni i potężnej liczbie punktów wykorzystywany jest tzw. trick jądrowy (*kernel trick*), dzięki któremu możliwości obliczeniowe są zależne od wymiaru przestrzeni pierwotnej. Wówczas iloczyn skalarny wektorów x_i x_j w przestrzeni zmiennych przekształconych jest równy tzw. funkcji jądrowej[114]:

$$\varphi(x_i) \cdot \varphi(x_j) = k(x_i x_j)\,. \qquad (2.41)$$

W literaturze dotyczącej rozpoznawania wzorców za pomocą metody SVM stosowane są następujące funkcje jądrowe[115]:

- liniowe

[113] J. Goszczyński, *Klasyfikacja tektur za pomocą SVM – maszyny wektorów wspierających,* „Inżynieria Rolnicza", 2006, nr 13, s. 121.
[114] A. Marcinkowska-Ochtyra, op. cit, s. 39.
[115] J. Goszczyński, op. cit., s. 121.

$$k(x_i x_j) = x_i^T x_j\,, \qquad (2.42)$$

- wielomianowe

$$k(x_i x_j) = (\gamma x_i^T x_j + r)^d,\ \ \gamma > 0, \qquad (2.43)$$

- radialne

$$k(x_i x_j) = exp(-\gamma \|x_i - x_j\|)^d,\ \ \gamma > 0, \qquad (2.44)$$

- sigmoidalne

$$k(x_i x_j) = \tanh(\gamma x_i^T x_j + r), \qquad (2.45)$$

gdzie:

γ – parametr gamma,

d – stopień wielomianu,

r – parametr przesunięcia (*bias*).

Dla każdej z funkcji jądrowych istnieje możliwość wyznaczenia tzw. parametru *C*, określanego kosztem kary (*cost of the penalty*), który jest odpowiedzialny za kontrolę kompromisu między błędami w klasyfikacji obserwacji a wymuszeniem marginesów. Wysokie wartości parametru stanowią o wysokiej karze nakładanej na obserwacje znajdujące się po złej stronie hiperpłaszczyzny lub wewnątrz marginesu. Z kolei, niskie wartości parametru są przyczyną mniejszej dokładności klasyfikacji, czyniąc klasyfikator mało skutecznym. W przypadku funkcji liniowej definiowany jest tylko parametr kary *C*, natomiast funkcja radialna i sigmoidalna wymagają dodatkowo zaimplementowania parametru *r*, który jest różny od zera i świadczy o optymalnym dopasowaniu danych treningowych, a także współczynnika gamma, mówiącego o szerokości funkcji Gaussa. Funkcja wielomianowa, oprócz wspomnianych parametrów wymaga również określenia stopnia wielomianu[116].

2.4 Sztuczne sieci neuronowe

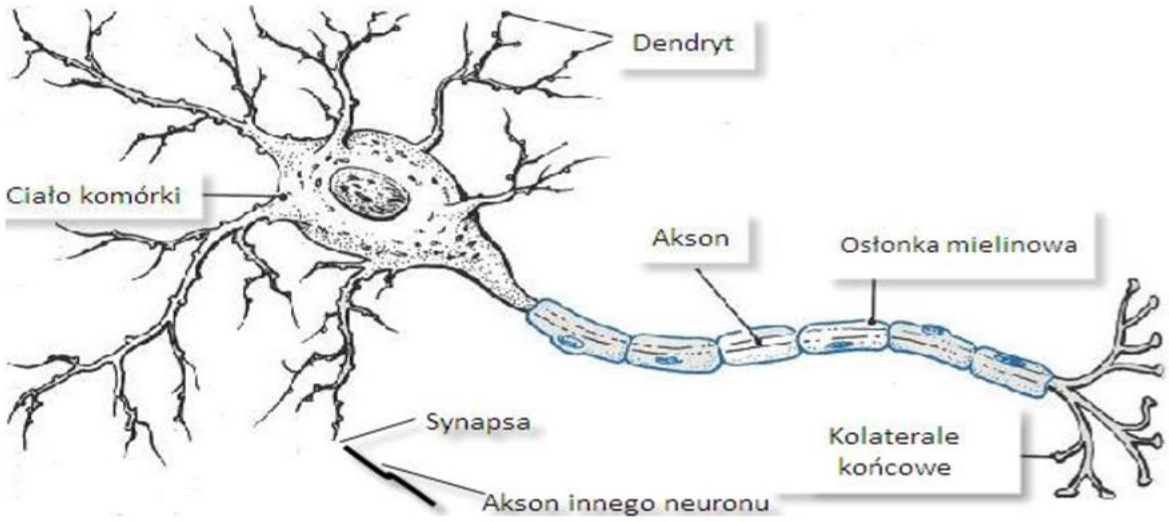
W naukach neurofizjologicznych sztuczne sieci neuronowe definiowane są jako systemy przetwarzania danych symulujące uczące się struktury mózgu. Z kolei, z matematycznego punktu widzenia, sztuczne sieci neuronowe to modele regresyjne

[116] A. Marcinkowska-Ochtyra, op. cit, s. 40.

zbudowane w oparciu o pewne klasy sparametryzowanych funkcji nieliniowych, których parametry estymuje się zwykle metodami nieparametrycznymi[117].

Źródeł idei powstania sztucznych sieci neuronowych należy doszukiwać się w badaniach nad budową i funkcjonowaniem ludzkiego mózgu, który jest pierwowzorem sieci neuronowych. Składa się on z 10^{10} komórek nerwowych, między którymi jest 10^{15} połączeń. Komórki nerwowe przetwarzają impulsy z częstotliwością 1 - 100 Hz. Daje to przybliżoną prędkość pracy mózgu 10^{18} operacji na sekundę, co wielokrotnie przewyższa możliwości obecnych superkomputerów[118]. Należy zaznaczyć, że sieć neuronowa jest bardzo uproszczonym modelem mózgu. Zbudowana jest ona z dużej liczby elementów przetwarzających informację zwanych neuronami, które połączone są ze sobą w pewien określony sposób. Model podstawowej biologicznej komórki nerwowej – neuronu, będącej częścią sieci neuronowej budującej ludzki mózg zilustrowano na rysunku 2.5.

Rys. 2.5. Model biologicznej komórki nerwowej



Źródło: A. Skrobała, *Zastosowanie sztucznych sieci neuronowych do optymalizacji rozkładów dawek w radioterapii stereotaktycznej obszarów wewnątrzczaszkowych*, Rozprawa doktorska, Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu, 2015, s. 20.

Zgodnie z działaniem komórki nerwowej, neuron jako podstawowa jednostka obliczeniowa w mózgu, otrzymuje sygnał wejściowy od swych dendrytów i tworzy sygnał wyjściowy przesyłając go wzdłuż jego aksonu. Akson jest połączony przez synapsy do dendrytów kolejnych neuronów, przetwarzając w ten sposób sygnał dalej. Moc z jaką każda

[117] P. Baster, K. Pocztowska, *Sieci neuronowe i polichotomiczne modele zmiennych jakościowych w analizie ryzyka kredytowego*, „Folia Oeconimca Cracoviensia", 2011, vol. LII, s. 6.
[118] Materiały edukacyjne AGH http://galaxy.uci.agh.edu.pl/~vlsi/AI/wstep/ (dostęp: 18.12.2019).

synapsa przetwarza sygnał jest zmienna, przyswajalna w trakcie uczenia. Wszystkie sygnały dostarczone przez dendryty do neuronu są sumowane w ciele komórki. Gdy suma jest większa od określonego progu, do aksonu wysyłany jest impuls elektryczny. Czyniąc założenie, że dokładny czas poszczególnych impulsów nie ma znaczenia oraz korzystając tylko z częstotliwości ich wytwarzania, można modelować ten sygnał statyczną funkcją aktywacji[119].

Historyczną pracą, w której po raz pierwszy zaprezentowano matematyczny opis komórki nerwowej i powiązano go z problemem przetwarzania danych była praca Warrena McCulloch`a i Waltera Pittsa, pt.: *A Logical Calculus of the Ideas Immanent in Nervous Activity*[120]. Autorzy opisali taką komórkę nerwową jako prostą bramkę logiczną tworzącą binarne wyjścia; do dendrytów neuronu dociera wiele sygnałów, które są integrowane w ciele komórki i, jeśli energia impulsu przekracza pewną wartość graniczną, zostaje wygenerowany sygnał wyjściowy przepuszczany przez akson[121]. Zauważono wówczas, że bardzo istotną własnością sztucznych sieci neuronowych jest ich zdolność do równoległego przetwarzania informacji, a za główną zaletę sieci uznano jej umiejętności uczenia się, co stanowi lepszą alternatywę w stosunku do tradycyjnego programowania[122].

Do podstawowych komponentów neuronów wykorzystywanych w sztucznych sieciach neuronowych zaliczamy[123]:

- wejścia (*inputs*), którymi docierają sygnały zewnętrzne x_i,
- wagi w_i (*weights*), których wartości stanowią względną ważność poszczególnych sygnałów wejściowych,
- funkcja aktywacji *act* (a*ctivation function*), określająca metodę obliczania pobudzenia e w oparciu o zestaw wag w_i i sygnałów wejściowych x_i,
- pobudzenie *e* (*excitation*) – wielkość skalarna determinująca aktywność neuronu,
- wartość wyjścia (*output*) neuronu *y*,

[119] K. Odrzywołek, *Wykorzystanie głębokich sieci neuronowych w weryfikacji mówcy Deep neural networks in speaker recognition*, Praca dyplomowa magisterska, Akademia Górniczo-hutnicza im. Stanisława Staszica w Krakowie, 2016, s. 15-16.
[120] W. S McCulloch, W. Pitts, *A Logical Calculus of the Ideas Immanent in Nervous Activity*, "The Bulletin of Mathemical Biophysics", 1943, vol. 5, no. 4, s. 115-133.
[121] S. Raschka, op. cit., s. 42.
[122] J. Siderska, *Analiza możliwości zastosowania sieci neuronowych do modelowania wartości kapitału społecznego w firmach IT*, „Economics and Management", 2013, nr 1, s. 86.
[123] K. Krawiec, J. Stefanowski, *Uczenie maszynowe i sieci neuronowe,* Wydanie II, Wydawnictwo Politechniki Poznańskiej, 2004, s. 83.

- funkcja przenosząca f (transfer function), której wartość warunkuje stan wyjścia neuronu *y* na podstawie jego aktywacji *e*.

Stosując powyższe symbole oraz oznaczając przez k liczbę wejść neuronu, jego funkcjonowanie można opisać wzorem[124]:

$$y = f\big(act(x_1, \dots, x_k, w_1, \dots, w_k)\big) \qquad (2.46)$$

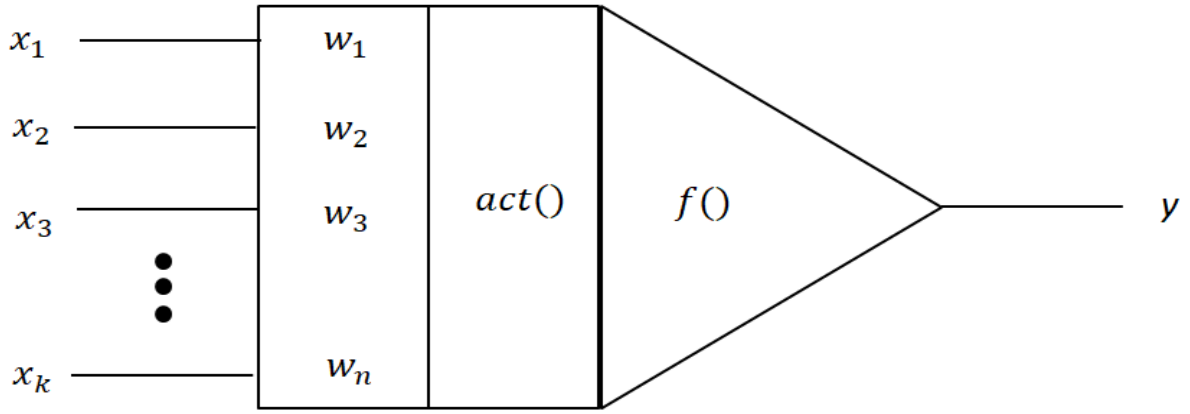
lub bardziej szczegółowo:

$$e = act(x_1, \dots, x_k, w_1, \dots, w_k), \qquad (2.47)$$

$$y = f(e). \qquad (2.48)$$

Tak opisany sztuczny neuron jest zwykle przedstawiany graficznie w postaci schematu ujętego na rysunku 2.6 lub w podobny sposób. Zbiegający się ku wyjściu trójkątny kształt ma tu odzwierciedlać podstawową funkcję neuronu, czyli agregację sygnałów wyjściowych.

Rys. 2.6. Podstawowy model sztucznego neuronu



Źródło: K. Krawiec, J. Stefanowski, *Uczenie maszynowe i sieci neuronowe*, Wydanie II, Wydawnictwo Politechniki Poznańskiej, 2004, s. 84.

W zapisie wektorowym model neuronu widoczny jest w równaniu (2.49):

$$y = f\big(act(x, w)\big). \qquad (2.49)$$

Poszczególne modele neuronów stosowane w różnych rodzajach sieci neuronowych różnią się zazwyczaj funkcją przenoszącą f i funkcją aktywacji *act*. W większości architektur funkcja aktywacji *act* wykonuje prostą sumę ważoną, czyli iloczyn skalarny wektora

[124] Ibidem, s. 84.

sygnałów wejściowych i wektora wag, pomniejszoną o stały parametr θ zwany progiem (*treshold*):

$$act(x, w) = \sum_{i=1}^{k} w_i x_i - \theta = w * x - \theta \ . \qquad (2.50)$$

Zmodyfikowane, przez określone wagi, sygnały wejściowe neuronu na skutek działania sumatora są sumowane dając w rezultacie pewien sygnał wewnętrzny, zwany łącznym pobudzeniem neuronu lub pobudzeniem postsynaptycznym. Ten sygnał może również być bezpośrednio przesyłany do wyjścia neuronu i traktowany jako sygnał wyjściowy.

W ciele neuronu do przetwarzania pobudzenia *e* wykorzystywana jest pewna funkcja przenosząca, zwana też funkcją wyjścia:

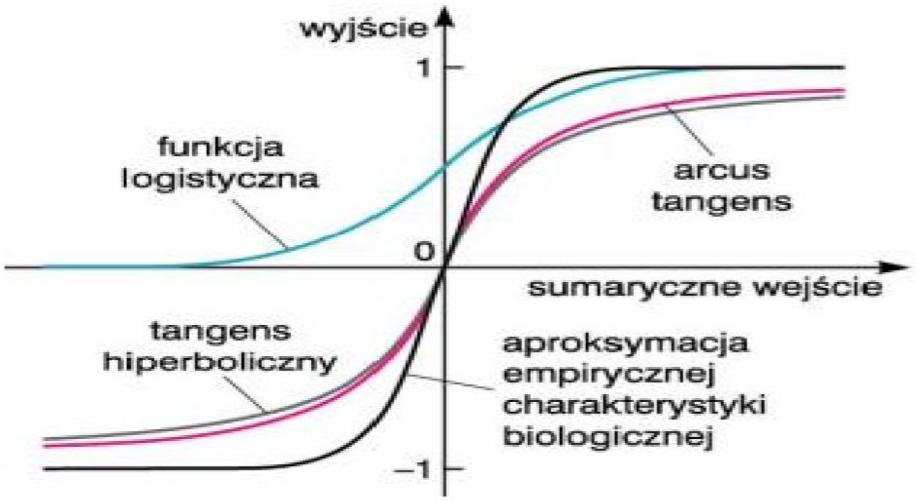
$$y = f(e) \ . \qquad (2.51)$$

Funkcja przenosząca f musi spełniać pewne warunki, aby zagwarantowane było prawidłowe działanie sieci oraz jej algorytm uczenia. Przede wszystkim funkcja f powinna być monotoniczna. Poza tym, ponieważ większość algorytmów optymalizujących jest oparta na metodach gradientowych posługuje się zazwyczaj f różniczkowalną w całej swej dziedzinie. Wyłącznie w nietypowych i rzadziej stosowanych w praktyce modelach neuronowych można czasami spotkać się z funkcjami przenoszącymi *f* nie spełniającymi tego ograniczenia, np. w przypadku zastosowania funkcji skoku jednostkowego w sieci Hopfielda[125].

Najczęściej spotykanymi funkcjami aktywacji neuronu *f(e)* jest funkcja liniowa (w których zachowania neuronów daje się łatwo objaśnić) lub funkcje nieliniowe. Dla funkcji nieliniowych suma iloczynów wag i sygnałów wejściowych jest przeliczana na wartości -1 albo 1 stosując odpowiednie funkcje matematyczne, najczęściej tzw. sigmoidalne, ale również, arcus tangens czy tangens hiperboliczny[126].

[125] K. Krawiec, J. Stefanowski, op. cit., s. 85-86.
[126] A. Skrobała, op. cit., s. 24.

Rys. 2.7. Funkcje aktywacji wiążące zsumowane wejścia neuronu z jego sygnałem wyjściowym



Źródło: A. Skrobała, *Zastosowanie sztucznych sieci neuronowych do optymalizacji rozkładów dawek w radioterapii stereotaktycznej obszarów wewnątrzczaszkowych*, Rozprawa doktorska, Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu, 2015, s. 24.

Proces uczenia się sieci neuronowej polega na pewnej optymalizacji wag sygnałów wejściowych. Architektura sieci jest ściśle związana z odpowiednią metodą doboru wag, tj. metodą uczenia. Uczenie takiej sieci odbywa się z „nauczycielem", to znaczy zbiorem o znanym rozwiązaniu. Metoda uczenia oparta jest na zasadzie minimalizacji funkcji kosztu reprezentującej błędy popełniane przez sieć[127].

Funkcję kosztu (*loss function*) opisującą miarę błędów popełnionych przez sieć dla zadań regresji liniowej można zapisać jako średnią kwadratów różnic (*mean square error*) między pożądanymi rezultatami (Z) a tymi wnioskowanymi przez sieć (Y)[128]:

$$\mathcal{L}(Y,Z) = \frac{1}{N}\sum_{n=1}^{N}(z_n - y_n)^2 \,. \qquad (2.52)$$

Z kolei, dla zadań klasyfikacji korzysta się ze średniej entropii krzyżowej (*cross entropy*):

[127] B. Darłak, M. Włodarczyk, *Zastosowanie sztucznej sieci neuronowej do uzupełnienia danych zbiornikowych*, „Przegląd Geologiczny", 2001, vol. 49, nr 9, s. 797-798.
[128] K. Odrzywołek, op. cit., s. 21.

$$\mathcal{L}(Y,Z) = -\frac{1}{N}\sum_{n=1}^{N}(z_n log y_n + (1-z_n)\log(1-y_n)). \qquad (2.53)$$

Oba przypadki mają swoje modyfikacje stosujące sumy w miejsce średnich.

Istnieje wiele metod uczenia sieci neuronowych. Do najważniejszych należy zaliczyć[129]:

- *Back Propagation* (metoda wstecznej propagacji błędów),
- *Quick Propagation* (metoda szybka propagacji błędów).
- *Quasi-Newton* (metoda zmiennej metryki),
- *Conjugate Gradients* (metoda gradientów sprzężonych),
- *Levenberg – Marquardt* (metoda paraboloidalnych modeli funkcji błędów),

Szczególny rodzaj sztucznych sieci neuronowych stanowią sieci neuronowe jednokierunkowe wielowarstwowe (*Multi Layer Perceptron* – MLP).

Sieć MLP to wielowarstwowa, jednokierunkowa sieć neuronowa, zbudowana z szeregu warstw neuronów, z których pierwszą warstwą (do której docierają sygnały wejściowe) nazywamy wejściową, ostatnią warstwę (do której przesyłany jest/są sygnały wejściowe) nazywamy wyjściową, a pośrednie warstwy (których może być wiele) nazywamy warstwami ukrytymi. Dzięki warstwowej budowie, model podejmowania decyzji w sieci MLP może być znacznie bardziej złożony i w rezultacie bardziej skuteczny. Neurony warstwy wejściowej, zamiast podejmować decyzje wynikowe, mogą rozpoznawać pewne elementarne cechy sygnału wejściowego. Analogicznie, następne warstwy sieci mogą obliczać coraz to bardziej złożone wielkości, i wówczas neurony warstwy wyjściowej mogą podejmować ostateczne decyzje na podstawie znacznie bardziej znaczących kategorii niż surowe zmienne wejściowe[130].

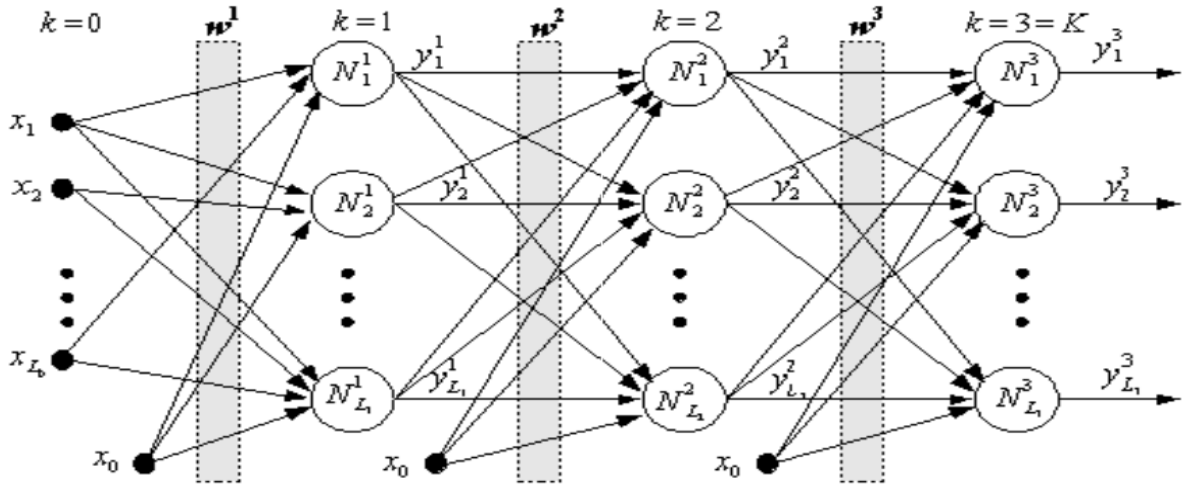
Perceptron wielowarstwowy o odpowiedniej strukturze jest w stanie modelować funkcję o prawie dowolnej złożoności. Liczba neuronów w sztucznych sieciach neuronowych waha się w granicach od kilkunastu do kilkuset. Sieci składające się z kilku tysięcy i więcej

[129]Materiały edukacyjne Katedry Inżynierii Komputerowej Uniwersytetu Rzeszowskiego, http://www.neurosoft.edu.pl/media/pdf/jbartman/sztuczna_inteligencja/NTI%20cwiczenie_5.pdf (dostęp: 07.01.2020).

[130] Materiały edukacyjne Katedry Cybernetyki i Robotyki Politechniki Wrocławskiej, https://kcir.pwr.edu.pl/~witold/ai/ml_nndeep_s.pdf (dostęp: 07.01.2020).

neuronów uważa się za duże. W teorii liczba warstw ukrytych w sieci neuronowej może być dowolna, lecz na podstawie analizy literatury przedmiotu, można stwierdzić, że do rozwiązywania praktycznych problemów, wystarczy jedna lub dwie warstwy ukryte, z tym, że w przypadku sieci wielowarstwowych jednokierunkowych nie można z góry określić odpowiedniej struktury sieci dla konkretnego zadania – należy ją ustalić eksperymentalnie[131].

Rys. 2.8. Schemat wielowarstwowej sieci perceptronowej posiadającej warstwę wejściową, dwie warstwy ukryte neuronów i warstwę wyjściową neuronów



Źródło: J. Protasiewicz, *Zastosowanie sieci neuronowych do analizy rynku energii elektrycznej w Polsce*, Rozprawa doktorska, Instytut Badań Systemowych Polskiej Akademii Nauk, Warszawa 2008, s. 43.

W sieci MLP przepływ informacji następuje tylko w jednym kierunku: od warstwy wejściowej do warstwy wyjściowej. Sygnały wejściowe[132]:

$$x(t) = [x_0(t), x_1(t), \dots, x_{L0}(t)]^T, \qquad (2.54)$$

gdzie: t = 1,2, ... jest dyskretnym czasem, a L_0- liczbą wejść w warstwie wejściowej, są przekazywane do pierwszej warstwy ukrytej neuronów. Neuron N_i^k (i-ty w warstwie k-tej) ma N_{k-1} wejść, zatem sygnał wejściowy neuronu N_i^k jest związany z sygnałem wyjściowym neuronów warstwy $k-1$ w następujący sposób:

[131] J. Dawidowicz, *Badania struktur sieci neuronowych typu MLP do oceny układu stref ciśnienia systemu dystrybucji wody*, „Civil and Environmental Engineering/Budownictwo i Inżynieria Środowiska", 2015, nr 6, s. 54.
[132] J. Protasiewicz, *Zastosowanie sieci neuronowych do analizy rynku energii elektrycznej w Polsce*, Rozprawa doktorska, Instytut Badań Systemowych Polskiej Akademii Nauk, Warszawa 2008, s. 43-44.

$$x_i^{(k)} = \begin{cases} x_i(t) & dla\ k = 1 \\ y_i^{(k-1)}(t) & dla\ k = 2, \dots, K \\ x_0^{(k)} & dla\ i = 0, k = 1, \dots, K \end{cases}, \qquad (2.55)$$

przy czym *k = 1, ..., K* oznacza kolejne warstwy, a *K* jest maksymalną liczbą warstw sieci.

Jeżeli wektor wag związanych z neuronem N_i^k zostanie zapisany jako:

$$w_i^k(t) = \left[w_{i,0}^k(t), w_{i,1}^k(t), \dots, w_{i,L_{k-1}}^k(t)\right]^2, \qquad (2.56)$$

gdzie: *k = 1, ...,K, i = 1, ...,*L_k, to wagi neuronów warstwy *k* tworzą macierz wag W^k. Sygnał wyjściowy neuronu N_i^k w chwili *t* − tej, dla *t = 1, 2, ...* jest określony jako:

$$e_i^{(k)}(t) = \sum_{j=0}^{L_{k-1}} w_{ij}^k(t)\, x_j^{(k)}(t), \qquad (2.57)$$

$$y_i^{(k)}(t) = f\left(e_k^{(k)}(t)\right), \qquad (2.58)$$

zaś sygnał wyjściowy całej *k* − tej warstwy sieci wynosi:

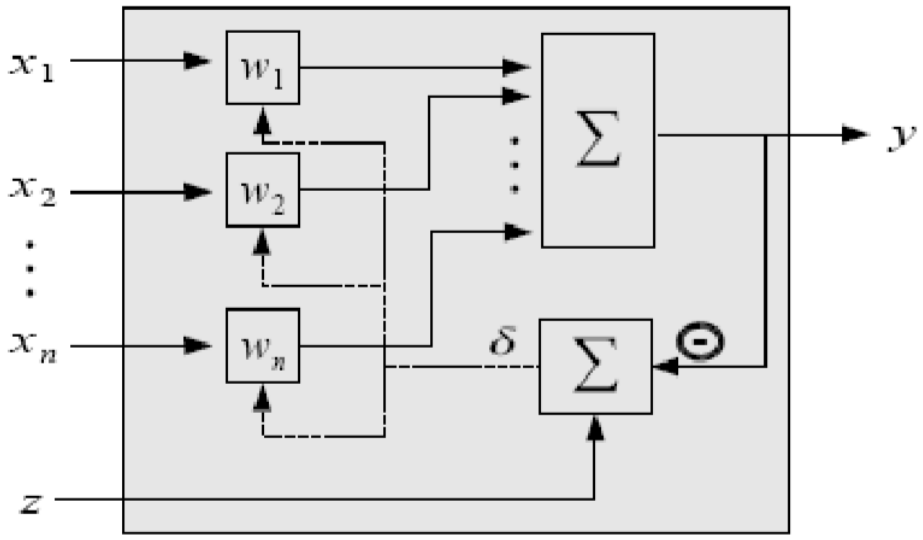
$$y^{(k)}(t) = f\left(x^{(k)}(t) \cdot W^{(k)}(t)\right), \qquad (2.59)$$

i przepływając przez kolejne warstwy sieci dociera do warstwy wyjściowej, stając się jednocześnie sygnałem wyjściowym całej sieci.

Przez wiele lat nie znano metody wydajnego uczenia sieci wielowarstwowych. Dopiero w połowie lat 80. XX wieku zaproponowany został algorytm wstecznej propagacji błędów (*backpropagation*), polegający na tym, że mając określony błąd $\delta_i^{(k)}$ formułujący się podczas realizacji *k*-tego kroku procesu uczenia w neuronie *i* można podawać ten błąd wstecz (w stronę przepływu informacji – stąd nazwa algorytmu) do wszystkich tych neuronów, których sygnały składały się na wejścia dla *i*- tego neuronu[133].

[133] Materiały edukacyjne Katedry Inżynierii Komputerowej Uniwersytetu Rzeszowskiego, *http://www.neurosoft.edu.pl/media/pdf/tkwater/sztuczna_inteligencja/2_alg_ucz_ssn.pdf* (dostęp: 11.01.2020).

Rys. 2.9. Schemat algorytmu wstecznej propagacji w sieci MPL

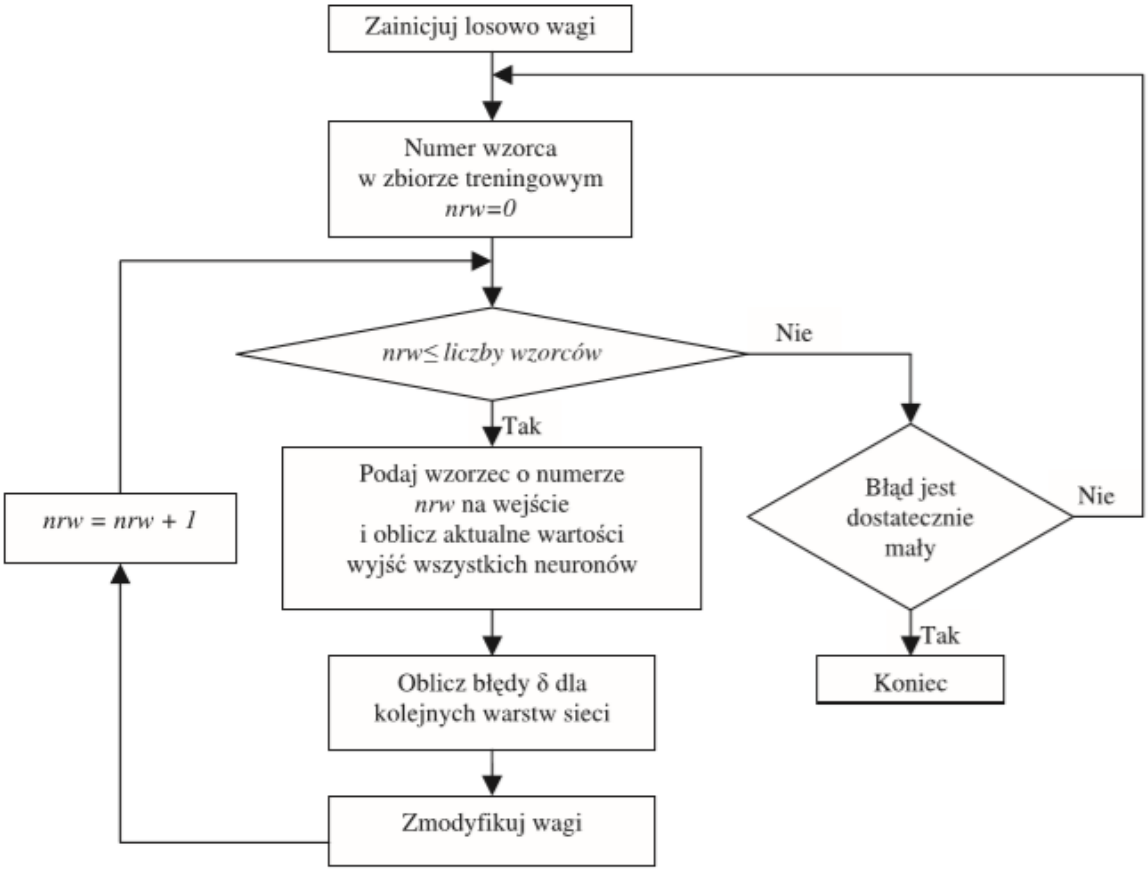


Źródło: A. Burda, *Prognozowanie kondycji ekonomiczno-finansowej przedsiębiorstw z wykorzystaniem sztucznych sieci neuronowych*, „Barometr Regionalny", 2006, nr 6, s. 70.

Proces uczenia sieci neuronowej został przedstawiony w sposób algorytmiczny na rys. 2.9 i 2.10. Rozpoczynamy od wygenerowania losowego wektora wag odpowiadającego całej sieci. Następnie poprzez kolejno dostarczane przykłady uczące, właściwy algorytm uczenia, dokonuje stopniowej zmiany wag, tak aby uzyskać oczekiwane wartości sygnałów na wyjściu sieci. Wektor sygnałów wejściowych *x(t)* przetwarzany jest na wartość wyjściową *y*. Następnie jest ona porównywana z oczekiwaną na wyjściu wartością *z*. Jeśli różnica obu sygnałów $\delta = 0$ oznacza to, że neuron prawidłowo spełnia swoje zadanie i korekta wag nie jest potrzebna. W przeciwnym wypadku wagi w_i należy optymalizować proporcjonalnie do wielkości błędu δ[134].

[134] A. Burda, *Prognozowanie kondycji ekonomiczno-finansowej przedsiębiorstw z wykorzystaniem sztucznych sieci neuronowych,* „Barometr Regionalny", 2006, nr 6, s. 69.

Rys. 2.10. Algorytm uczenia sieci neuronowej



Źródło: A. Burda, *Prognozowanie kondycji ekonomiczno-finansowej przedsiębiorstw z wykorzystaniem sztucznych sieci neuronowych*, „Barometr Regionalny", 2006, Nr. 6, s. 70.

Celem procesu uczenia sieci neuronowej jest doprowadzenie do możliwie pełnej zbieżności pomiędzy $y^{(k)}$(t) a $z^{(k)}(t)$, czyli do minimalizacji funkcji kosztu (2.60), gdzie L_k jest liczbą neuronów wyjściowych[135]:

$$Q = \frac{1}{2}\sum_{j=1}^{L_k}(z^{(k)}(t) - y^{(k)}(t))^2 \quad . \qquad (2.60)$$

Jedną z możliwości optymalizacji stanowi wykorzystanie metody gradientowej. Zatem dla dowolnego neuronu wielkość korekty dowolnej jego wagi można opisać wzorem:

$$\Delta W_{ij}{}^{(k)} = -\eta \frac{\partial Q(t)}{\partial W_{ij}{}^{(k)}(t)} \quad . \qquad (2.61)$$

[135] Ibidem, s. 70.

Współczynnik uczenia η stanowi o długości kroku minimalizacji w kierunku określonym przez gradient. Krok zbyt duży będzie powodować, że algorytm może nie znaleźć minimum, z kolei krok zbyt mały będzie skutkował długim czasem schodzenia w kierunku minimum funkcji błędu. Istnieje kilka technik wyznaczania optymalnej wartości tego parametru uczenia. Najprostsze jest przyjęcie współczynnika o stałej wartości dla całego procesu uczenia formułowanego do każdego problemu eksperymentalnie. Najczęściej przyjmuje się, że współczynnik uczenia powinien przybierać wartości z przedziału $\eta \in < 0.001, 10 >$ [136].

Ponieważ Q jest zależny od y, który jest funkcją wektora wag W, więc prawą stronę równania (2.61) łatwo można przekształcić do postaci:

$$\frac{\partial Q(t)}{\partial W_{ij}^{(k)}(t)} = \frac{\partial Q(t)}{\partial y_i^{(k)}(t)} \frac{\partial y_i^{(k)}(t)}{\partial W_{ij}^{(k)}(t)} . \qquad (2.62)$$

Skoro zgodnie z równaniem (2.60):

$$Q = \frac{1}{2}\sum_{j=1}^{L_k} (z^{(k)}(t) - y^{(k)}(t))^2 , \qquad (2.60)$$

to:

$$\frac{\partial Q(t)}{\partial y_i^{(k)}(t)} = -\left(z^{(k)}(t) - y^{(k)}(t)\right) = \delta_i^{(k)}(t). \qquad (2.63)$$

Łatwo też zauważyć zależność (2.64) we wzorze (2.62):

$$\frac{\partial y_i^{(k)}(t)}{\partial W_{ij}^{(k)}(t)} = x_j^{(k)}(t). \qquad (2.64)$$

Ostatecznie wielkość korekty dowolnej wagi, dowolnego neuronu w sieci, przy dodaniu j-tego wzorca uczącego można przedstawić za pomocą wzoru (2.65):

$$\Delta W_i{}^j = -\eta \delta_i^{(k)}(t) x_j^{(k)}(t) . \qquad (2.65)$$

[136] J. Protasiewicz, op. cit., s. 49.

Istnieją różne kryteria wyznaczające moment zakończenia nauki sieci, np.[137]:

- Nauka jest przerywana, gdy zmiana całościowej funkcji kosztu jest mniejsza od zadanego parametru: $|\Delta Q| < \epsilon$.
- Zakończenie uczenia sieci ma miejsce, gdy wartość funkcji kosztu dla zbioru uczącego osiągnie wartość mniejszą od wcześniej zadanego parametru: $|Q| < \epsilon$.
- Nauka nie jest dalej prowadzona, gdy spada zdolność sieci do generalizacji. Można wyodrębnić ze zbioru testowego podzbiór walidacyjny i badać na nim błąd klasyfikacji w trakcie nauki sieci. Trening przerywa się w sytuacji, gdy wraz ze spadkiem funkcji kosztu, błąd zbioru walidacyjnego zaczyna wzrastać. Strategia ta ma zapobiec nadmiernemu przeuczeniu sieci.
- Naukę sieci przerywa się po określonej liczbie epok.

[137] H. Jurkiewicz, *Nieeuklidesowe sieci neuronowe*, Praca magisterska, Uniwersytet Mikołaja Kopernika Wydział Fizyki, Astronomii i Informatyki Stosowanej Katedra Informatyki Stosowanej, Toruń 2009, s. 18-19.

Rozdział 3. Metody strojenia i oceny skuteczności modelu

3.1 Problem nadmiernego dopasowania modelu

Nadmierne dopasowanie to tendencja metod eksploracji danych do dostosowywania modelu do danych kosztem uogólniania nieznanych wcześniej punktów danych[138].

Nadmierne dopasowanie jest cechą modelu, którego złożoność określona liczbą stopni swobody jest taka, że może on precyzyjnie dopasować się do elementów zbioru uczącego, zwłaszcza jeżeli są one obarczone szumem. Nadmierne dopasowanie jest zjawiskiem lokalnym, tj. w pewnych zakresach zmiennych wejściowych funkcja modelująca wykorzystuje pewną liczbę stopni swobody w sposób, który skutkuje dokładnym dopasowaniem się modelu do pewnych elementów zbioru uczącego[139].

Algorytm uczący trenowany jest przeważnie na pewnym zbiorze obserwacji (zbiór uczący), dla których znane są prawidłowe wyniki. Zakłada się, że po nauczeniu algorytmu można zastosować go do predykcji wyników również dla innych obserwacji, czyli algorytm w procesie uczenia uogólni zaobserwowane prawidłowości w zbiorze uczącym, tworząc regułę decyzyjną rozpoznającą wszelkie podobne obserwacje. Jednakże w sytuacji, gdy uczenie jest zbyt długie i skomplikowane, lub gdy przypadki uczące są nieliczne, model może "wymyślić" prawidłowości, które w rzeczywistości nie występują, a są efektem przypadkowych błędów w zbiorze danych uczących. W rezultacie tego ‘przeuczenia’ spada jakość algorytmu użytego do innych danych niż te, na których się uczył, choć dla obserwacji w zbiorze uczącym jest coraz lepszy.

Podsumowując, zjawisko nadmiernego dopasowania występuje, gdy model jest zbyt skomplikowany w porównaniu do liczby lub zaszumienia danych uczących. Wobec takich sytuacji możliwe są następujące rozwiązania[140]:

- uproszczenie modelu poprzez dobór modelu zawierającego mniej parametrów (np. modelu liniowego zamiast modelu wielomianowego), redukcja liczby atrybutów w danych uczących oraz ograniczenie modelu (regularyzacja),

[138] F. Provost, T. Fewcett, *Analiza danych w biznesie. Sztuka podejmowania skutecznych decyzji*, Wydawnictwo Helion, Gliwice 2015, s. 122.
[139] S. Jankowski, *Statystyczne systemy uczące się – modelowanie i klasyfikacja*, Materiały do wykładu i projektu Sieci neuronowe i neurokomputery, 2012.
[140] A. Geron, op. cit., s. 46.

- zdobycie większej ilości danych uczących,
- pozbycie się zaszumienia w danych uczących (np. na skutek usunięcia elementów odstających lub błędnych danych).

Jak łatwo się domyślić algorytm uczący może ulec też zjawisku przeciwnemu do przetrenowania, czyli niedotrenowaniu (niedopasowaniu do modelu). Występuje ono wtedy, gdy model jest zbyt prosty, aby wyuczyć się struktur danych uczących.

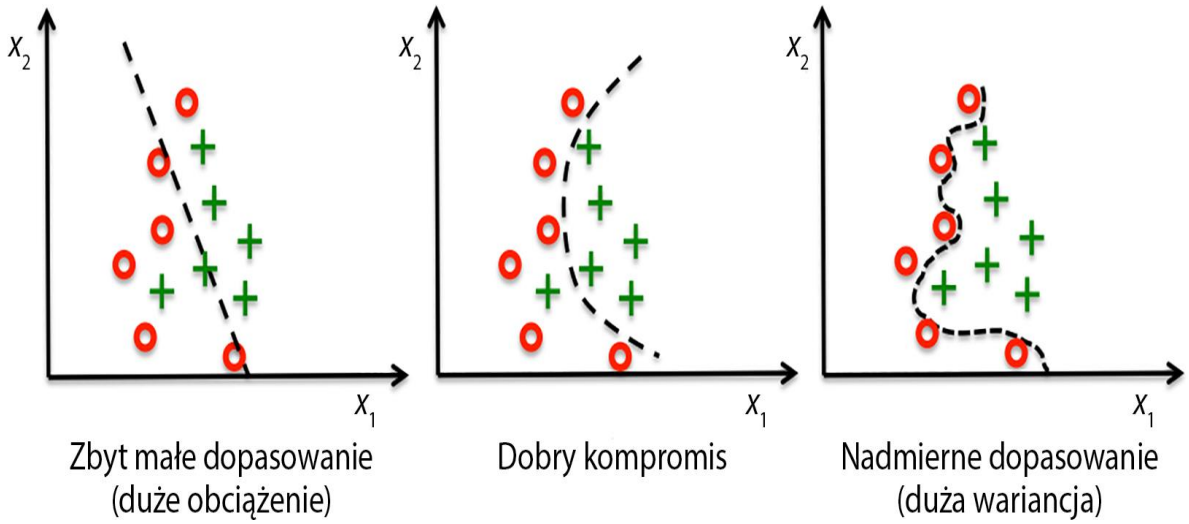
Głównymi sposobami rozwiązania tego problemu są[141]:

- wybór bardziej złożonego modelu wykorzystującego większą liczbę parametrów,
- zmniejszenie ograniczeń modelu (np. zredukowanie hiperparametru regularyzacji),
- zwiększenie liczby cech algorytmu uczącego (inżynieria cech).

3.2 Regularyzacja modelu

Model cechujący się przetrenowaniem bywa również nazywany modelem o dużej wariancji, natomiast model charakteryzujący się zbyt małym dopasowaniem definiowany jest także jako model o dużym obciążeniu.

Rys. 3.1. Przykłady nadmiernego i zbyt małego dopasowania



Źródło: S. Raschka, *Python. Uczenie maszynowe*, Wydawnictwo Helion, Gliwice 2018, s. 84.

[141] Ibidem, s. 48.

Jedną z metod znalezienia dobrego kompromisu pomiędzy wariancją a obciążeniem jest ustalenie optymalnej złożoności modelu poprzez jego regularyzację. Jest to bardzo dobry sposób na kontrolowanie współliniowości (dużej wzajemnej korelacji cech), odfiltrowanie szumów w danych oraz zapobieganie przetrenowaniu. Istotą regularyzacji jest dołączenie do modelu dodatkowych danych (obciążenia), karcących bardzo duże wartości wag[142].

Typowe podejście do regularyzacji estymacji polega na dodaniu do minimalizowanej funkcji celu Q, czynnika kary Ω zależnego od wektora parametrów θ. W efekcie zregularyzowana funkcja kryterialna ma następującą postać[143]:

$$Q_R(\theta) = Q(\theta) + k\Omega(\theta), \qquad (3.1)$$

gdzie k jest skalarnym współczynnikiem regularyzacji.

Plusem regularyzacji jest poprawa identyfikowalności modelu bez wymogu modyfikacji jego struktury, a minusem wzrost złożoności obliczeniowej i zwiększenie obciążenia estymatorów.

Minimalizując $Q_R(\theta)$ otrzymuje się obciążony estymator:

$$\hat{\theta}_R = arg\min_{\theta} Q_R(\theta), \qquad (3.2)$$

którego obciążenie b_R można wyliczyć linearyzując Q_R poprzez rozwinięcie w szereg Taylora:

$$b_R = -k(H_v + kH_\Omega)^{-1}\nabla_\theta\Omega\,, \qquad (3.3)$$

gdzie macierze hessianu H i gradient ∇ są określone następująco:

$$H_v = \frac{\partial^2 Q}{\partial\theta^T\partial\theta}\,, \qquad (3.4)$$

$$H_\Omega = \frac{\partial^2 \Omega}{\partial\theta^T\partial\theta}\,, \qquad (3.5)$$

$$\nabla_\theta\Omega = \frac{\partial\Omega}{\partial\theta}\,. \qquad (3.6)$$

[142] S. Raschka, op. cit., s. 84.
[143] A. J. Polak, J. Mroczka, *Regularyzacja identyfikacji obiektów złożonych opisanych modelami nieliniowymi*, „Pomiary Automatyka Kontrola", 2007, vol. 53, nr 9, s. 191.

Obciążenie narastające wraz ze współczynnikiem k jest kosztem za uzyskiwane wówczas zmniejszanie się wariancji S_R. Macierz kowariancji określona jest ogólnym wzorem:

$$S_R = (H_v + kH_\Omega)^{-1} P (H_v + kH_\Omega)^{-1}\,, \qquad (3.7)$$

$$P = E\left(\nabla_\theta Q \cdot \nabla_\theta^T Q\right). \qquad (3.8)$$

Najczęściej stosowaną karą w technice regularyzacji jest suma kwadratów wag, często nazywana „normą L2" dla *w*. Powód jej stosowania ma zwykle charakter techniczny, ale zasadniczo funkcje mogą bardziej dopasowywać się do danych, gdy pozwolimy im na posiadanie olbrzymich dodatnich i ujemnych wag. Suma kwadratów wag dostarcza większą „karę", gdy wagi mają duże wartości bezwzględne[144]. Wówczas składnik regularyzacyjny przyjmuje postać:

$$\Omega(\theta) = \frac{1}{2}\|w\|_2^2\,, \qquad (3.9)$$

$$Q_R(\theta) = Q(\theta) + k\frac{1}{2}\|w\|^2. \qquad (3.10)$$

Stopień regularyzacji przeprowadzonej na etapie nauki algorytmu można kontrolować również za pomocą hiperparametrów, zwanych także parametrami strojenia. Najprościej mówiąc hiperparametrami nazywamy parametry algorytmu uczącego (nie całego modelu) używane do sterowania zachowaniem modelu. Nie są one modyfikowane przez sam algorytm uczący; należy je wyznaczyć tuż przed rozpoczęciem procesu uczenia i w jego trakcie ich wartości pozostają niezmienne. Do hiperparametrów zaliczamy m. in. parametr kary *C* (stosowany w algorytmie SVM), stopień wysokości drzewa decyzyjnego, minimalną liczbę próbek w węźle, czy inne parametry ograniczające kształt drzewa.

Do optymalnego wyboru hiperparametrów modelu możemy posłużyć się metodą przeszukiwania siatki. Koncepcja kryjąca się za tą techniką jest bardzo prosta – mamy do czynienia z wyczerpującą techniką wyszukiwania, w której umieszczamy listę wartości różnych parametrów, a program ocenia skuteczność modelowania dla każdej kombinacji tych wartości parametrów w celu otrzymania optymalnej konfiguracji[145].

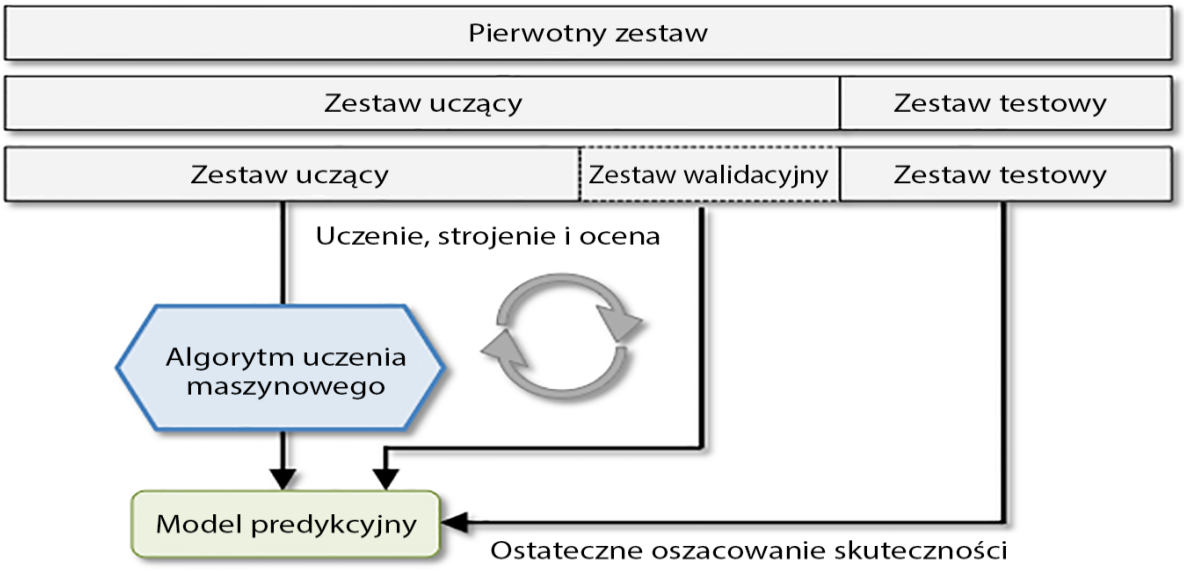
[144] F. Provost, T. Fewcett, op. cit., s. 144.
[145] S. Raschka, op. cit., s. 192.

3.3 Metoda wydzielania

Klasycznym i bardzo popularnym sposobem szacowania skuteczności uogólniania wyuczonych reguł przez model jest metoda wydzielenia. Polega ona na podzieleniu danych na trzy zestawy: zbiór uczący, zbiór walidacyjny i zbiór testowy. Zbiór danych uczących stosowany jest do trenowania różnych modeli, których skuteczność jest potem sprawdzana na próbkach walidacyjnych. Zaletą wydzielenia zbioru testowego z próbek nieznanych modelowi w fazie uczenia i doboru modelu jest uzyskanie bardziej niezależnego oszacowania zdolności uogólniania klasyfikacji. Na rysunku 3.2 przedstawiony został ogólny schemat metody wydzielania, w której zbiór danych walidacyjnych jest wykorzystywany do oceny skuteczności wytrenowanego modelu przy zastosowaniu różnych wartości hiperparametrów. Gdy uznamy dostrojenie wartości hiperparametrów za optymalne, możemy przejść do oszacowania skuteczności modelu względem danych testowych[146].

Uzasadnieniem takiego podejścia jest założenie, że nawet jeśli algorytm uczenia maszynowego może podlegać zjawisku przeuczenia na skutek istnienia losowych błędów czy artefaktów w zestawie danych uczących, to jest mało prawdopodobne, że niezależny zestaw walidacyjny składać się będzie z takich samych losowych fluktuacji. Dlatego wydzielenie odpowiednio dużego zestawu walidacyjnego pozwala na bezpieczniejszą realizację procesu regularyzacji modelu[147].

Rys. 3.2. Schemat metody wydzielania



Źródło: S. Raschka, *Python. Uczenie maszynowe*, Wydawnictwo Helion, Gliwice 2018, s. 181.

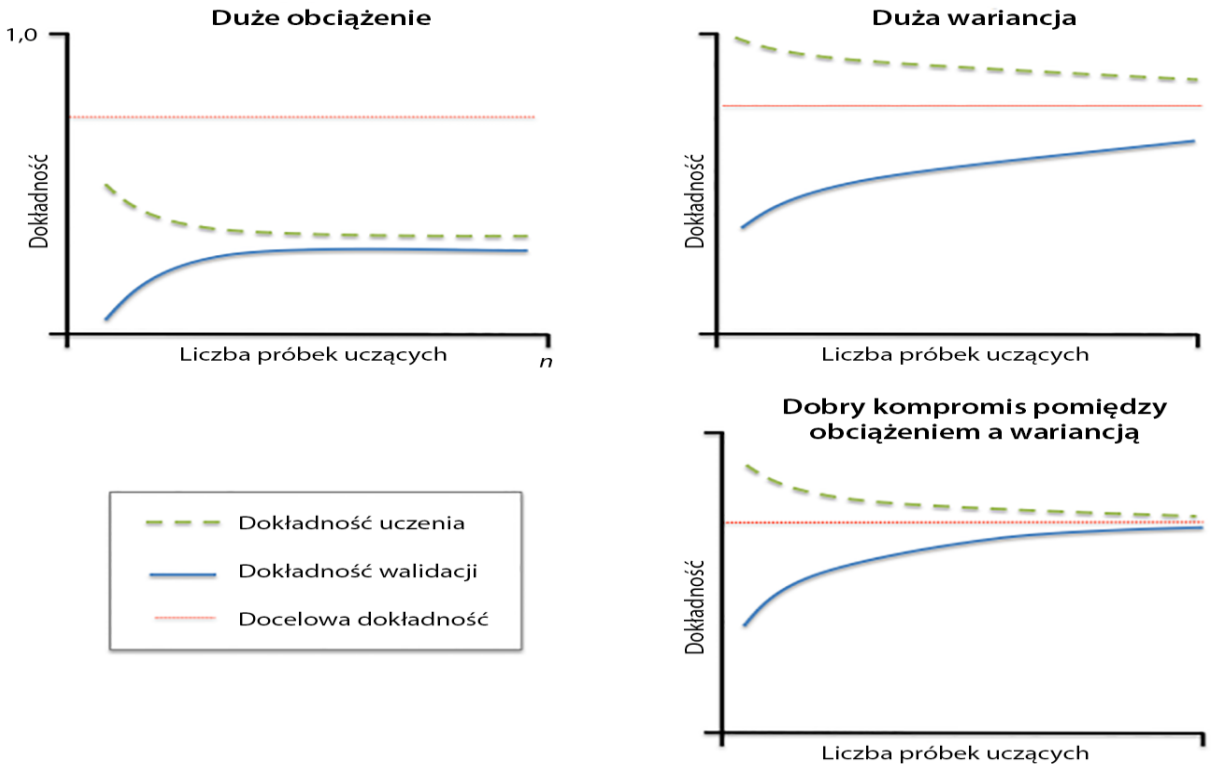
[146] Ibidem, s. 181.
[147] K. Krawiec, J. Stefanowski, op. cit., s. 45.

Jedną z częściej wymienianych wad metody wydzielania jest duża wrażliwość oszacowania na sposób podziału danych. Innymi słowy, wyniki mogą się istotnie różnić w zależności od wielkości poszczególnych podzbiorów.

Do zobrazowania skuteczności generalizacji algorytmu uczącego możemy wykorzystać krzywe uczenia i krzywe walidacji. Jak widać na rysunku 3.3 model wykazuje nadmierną wariancję, gdy krzywe walidacji oraz krzywe uczenia oddalają się od siebie (zauważalna jest duża przerwa między nimi), natomiast w sytuacji, gdy krzywe schodzą do siebie, ale znajdują się poniżej wymaganego poziomu dokładności, to w modelu występuje duże obciążenie. Celem jest znalezienie dobrego kompromisu pomiędzy obciążeniem a wariancją.

Poprzez utworzenie wykresu dokładności uczenia i walidacji modelu jako funkcji rozmiaru zbioru uczącego, bez trudu możemy się dowiedzieć, czy dodanie większej liczby obserwacji do modelu pomoże rozwiązać problem z dopasowaniem modelu. Krzywe uczenia i walidacji możemy zastosować również do oceny skuteczności strojenia hiperparametrów, wówczas wykres dokładności uczenia i walidacji tworzony jest wobec wartości parametrów algorytmu uczącego, np. wysokości drzewa decyzyjnego.

Rys. 3.3. Graficzne przedstawienie za pomocą krzywych uczenia i walidacji przetrenowania, niedostatecznego dopasowania modelu oraz właściwego kompromisu pomiędzy obciążeniem a wariancją



Źródło: S. Raschka, *Python. Uczenie maszynowe*, Wydawnictwo Helion, Gliwice 2018, s. 187.

3.4 K-krotny sprawdzian krzyżowy

Teoria sprawdzianu krzyżowego została zapoczątkowana przez amerykańskiego statystyka Seymoura Geissera[148]. Pozwala ona chronić się przed tzw. błędem trzeciego rodzaju i skutecznie ocenić trafność prognostyczną modelu. Bez jej zastosowania nie można mieć pewności, czy model będzie w stanie odpowiednio klasyfikować dane, które nie były wykorzystywane do jego konstruowania[149].

Sprawdzian krzyżowy jest bardziej zaawansowaną techniką uczenia i testowania danych wydzielonych. Przy jego zastosowaniu chcemy poznać nie tylko proste oszacowanie skuteczności uogólniania modelu, ale również pewne statystyki odnoszące się do szacowanej skuteczności. Kluczowe mogłoby być dla nas poznanie średniej i wariancji skuteczności generalizacji modelu, aby móc zrozumieć, jakie będzie przewidywane zróżnicowanie skuteczności w różnych zestawach danych. Ta wariancja ma istotne znaczenie przy ocenie zaufania do szacowanej skuteczności[150].

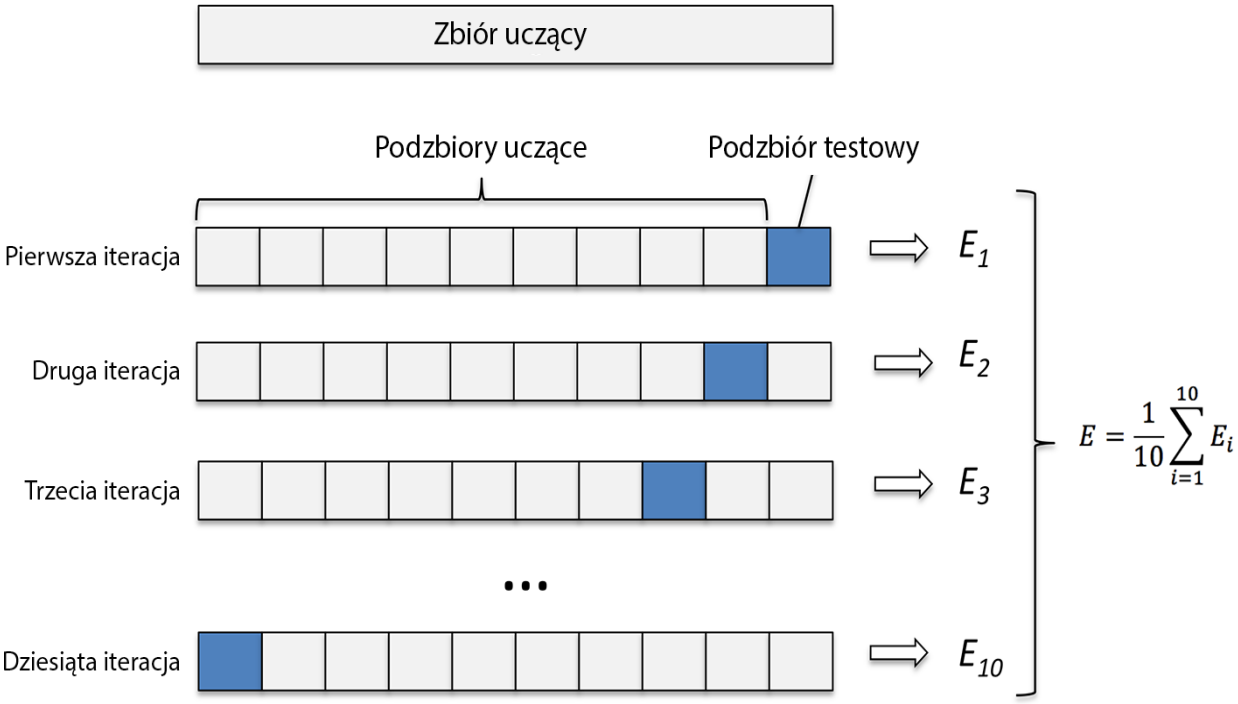
W k-krotnym sprawdzianie krzyżowym w sposób losowy rozdzielamy zestaw danych uczących na *k* niezależnych podzbiorów, gdzie *k-1* podzbiorów jest wykorzystywanych do trenowania modelu i strojenia parametrów, a tylko jeden do jego testowania. Mechanizm ten jest powtarzany k-krotnie, dzięki czemu uzyskujemy *x* modeli i oszacowań skuteczności. Wyliczamy następnie średnią skuteczność modeli w oparciu o różne i niezależne od siebie podzbiory, przez co w konsekwencji otrzymujemy oszacowanie skuteczności, które jest mniej czułe na podział danych od metody wydzielenia[151].

K-krotna walidacja krzyżowa dokonuje próbkowania bez zwracania. W takim przypadku, każda próbka znajdzie się tylko raz w zestawie treningowym lub testowym, na skutek czego szacowanie skuteczności modelu będzie związane z mniejszą wariancją niż w metodzie wydzielania[152].

---

[148] S. Geisser, *The Predictive Sample Reuse Method with Applications*,"Journal of the American Statistical Association", 1975, vol. 70.
[149] Encyklopedia internetowa, http://encyklopedia.naukowy.pl/Walidacja_krzy%C5%BCowa (dostęp: 13.02.2020).
[150] F. Provost, T. Fawcett, op. cit, s. 134.
[151] S. Raschka, op. cit., s. 182.
[152] Strona internetowa Machine Learning i Data Science, https://mlinpl.pl/walidacja-krzyzowa-kroswalidacja/ (dostęp: 13.02.2020).

Rys. 3.4. Schemat k-krotnego sprawdzianu krzyżowego dla 10 wydzielonych podzbiorów



Źródło: S. Raschka, *Python. Uczenie maszynowe*, Wydawnictwo Helion, Gliwice 2018, s. 181.

Rysunek 3.4 ilustruje 10-krotną walidację krzyżową w ramach, której zbiór uczący został podzielony na 10 podzbiorów. W przykładzie tym zgodnie z procedurą zostanie wykonanych 10 iteracji. W każdej z iteracji 9 podzbiorów zostanie wykorzystanych do uczenia modelu, a jeden do testowania. Następnie wyniki uzyskane w każdej z iteracji są uśredniane w celu otrzymania finalnej skuteczności modelu. Domyślna wartość parametru k wynosi 10 i jest to zarazem najczęściej stosowana wartość. W przypadku bardzo dużych zbiorów danych możliwe jest zmniejszenie wartości parametru k (np. k=5) i wciąż uzyskiwanie dokładnych oszacowań skuteczności modelu przy jednoczesnym ograniczeniu kosztów przetwarzania danych.

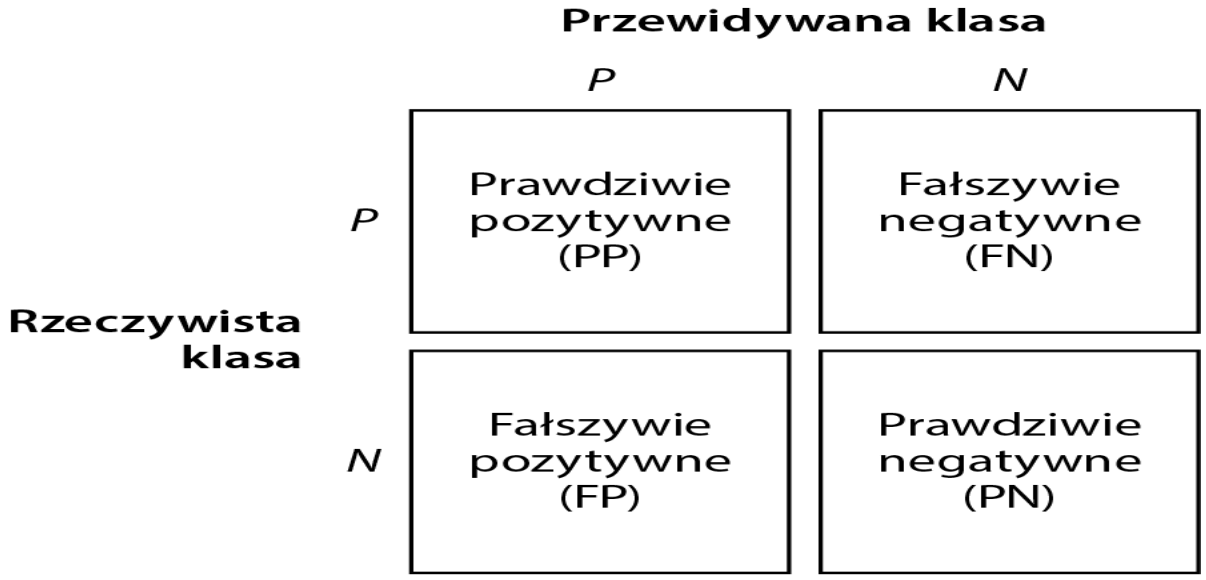
Warto zauważyć, że powiązanie k-krotnego sprawdzianu krzyżowego z techniką przeszukiwania siatki daje bardzo dobre efekty w dostrajaniu skuteczności modelu uczenia przez zróżnicowanie jego wartości hiperparametrycznych.

3.5 Macierz pomyłek

Macierz pomyłek jest podstawowym narzędziem stosowanym do oceny jakości klasyfikacji. Dla problemu n klas stanowi ona macierz $n \times n$, której kolumny etykietowane są rzeczywistymi klasami, a wiersze klasami przewidywanymi. Każdy przykład w zestawie testowym ma etykietę klasy rzeczywistej oraz etykietę klasy prognozowanej przez

klasyfikator, których kombinacja przyporządkowuje, do której komórki macierzy dane wystąpienie się zalicza. W przypadku klasyfikacji binarnej macierz składa się z dwóch wierszy i dwóch kolumn[153].

Rys. 3.5. Macierz pomyłek



Źródło: S. Raschka, *Python. Uczenie maszynowe*, Wydawnictwo Helion, Gliwice 2018, s. 181.

Symbole PP, FN, FP, PN oznaczają cztery możliwe przypadki klasyfikacji binarnej. Jeśli przyjmiemy, że występują dwie klasy, negatywna i pozytywna, to znaczenie symboli jest następujące[154]:

- PP (prawdziwie pozytywne) – liczba pozytywnych przypadków poprawnie sklasyfikowanych jako należących do klasy pozytywnej;
- FN (fałszywie negatywne) – liczba pozytywnych przypadków niepoprawnie sklasyfikowanych jako należące do klasy negatywnej; błędy typu FN nazywa się błędami drugiego rodzaju;
- FP (fałszywie pozytywne) – liczba negatywnych przypadków niepoprawnie sklasyfikowanych jako należące do klasy pozytywnej; są to błędy pierwszego rodzaju;
- PN (prawdziwie negatywne) – liczba negatywnych przypadków poprawnie sklasyfikowanych jako należące do klasy negatywnej.

[153] F. Provost, T. Fewcett, op. cit., s. 189.
[154] M. Szeliga, *Data science i uczenie* maszynowe, Wydawnictwo Naukowe PWN, Warszawa 2017, s. 300.

Z macierzy pomyłek można bezpośrednio wyznaczyć szereg miar oceniających jakość klasyfikatora. Dwie najczęściej wykorzystywane miary jakości reguły decyzyjnej to dokładność (*accuracy* – ACC) oraz błąd klasyfikacji (*missclassification/error rate* – ERR). Miary te dostarczają informacji na temat liczby prawidłowo i nieprawidłowo skalsyfikowanych próbek, tj.:

$$ACC = \frac{PP + PN}{FP + FN + PP + PN} = 1 - ERR\,, \qquad (3.11)$$

$$ERR = \frac{FP + FN}{FP + FN + PP + PN}\,. \qquad (3.12)$$

Do pozostałych miar jakości klasyfikatorów zaliczamy miary precyzji (PPP i PPN) oraz zasięgu (czułość i specyficzność)[155]:

- Precyzja przewidywania pozytywnego, określona wzorem:

$$PPP = \frac{PP}{PP + FP}\,, \qquad (3.13)$$

mierzy proporcję prawidłowych pozytywnych klasyfikacji wobec wszystkich pozytywnych klasyfikacji.

- Precyzja przewidywania negatywnego, określona wzorem:

$$PPN = \frac{PP}{PN + FN}\,, \qquad (3.14)$$

wskazuje, z jaką pewnością możemy ufać przewidywaniom negatywnym, tzn. w jakim procencie przewidywania negatywne zgadzają się ze stanem faktycznie negatywnym.

- Czułość, nazywana też zwrotem, wyrażona wzorem:

$$czułość = \frac{PP}{PP + FN}\,, \qquad (3.15)$$

mierzy proporcję poprawnych klasyfikacji w stosunku wszystkich pozytywnych przypadków.

- Specyficzność, nazywana też swoistością, określona wzorem:

$$specyficzność = \frac{PP}{PP + FP}\,, \qquad (3.16)$$

[155] Ibidem, s. 301-302.

mierzy proporcję prawidłowych negatywnych klasyfikacji wobec wszystkich negatywnych przypadków.

- F-miara (F-score) jest średnią harmoniczną precyzji i czułości, określona jest wzorem:

$$F - miara = 2 \cdot \frac{precyzja \cdot czułość}{precyzja + czułość}. \qquad (3.17)$$

Do obliczenia F-miary wykorzystuje się średnią harmoniczną, a nie arytmetyczną, bowiem precyzja i czułość są wartościami względnymi i przyjmują wartości z przedziału od 0 do 1. W związku z tym zakres wartości F-miary również wynosi [0, 1], gdzie 1 oznacza model idealny. Wzrost precyzji modelu eksploracji danych powoduje spadek miary czułości i odwrotnie, wraz ze wzrostem czułości maleje precyzja modelu.

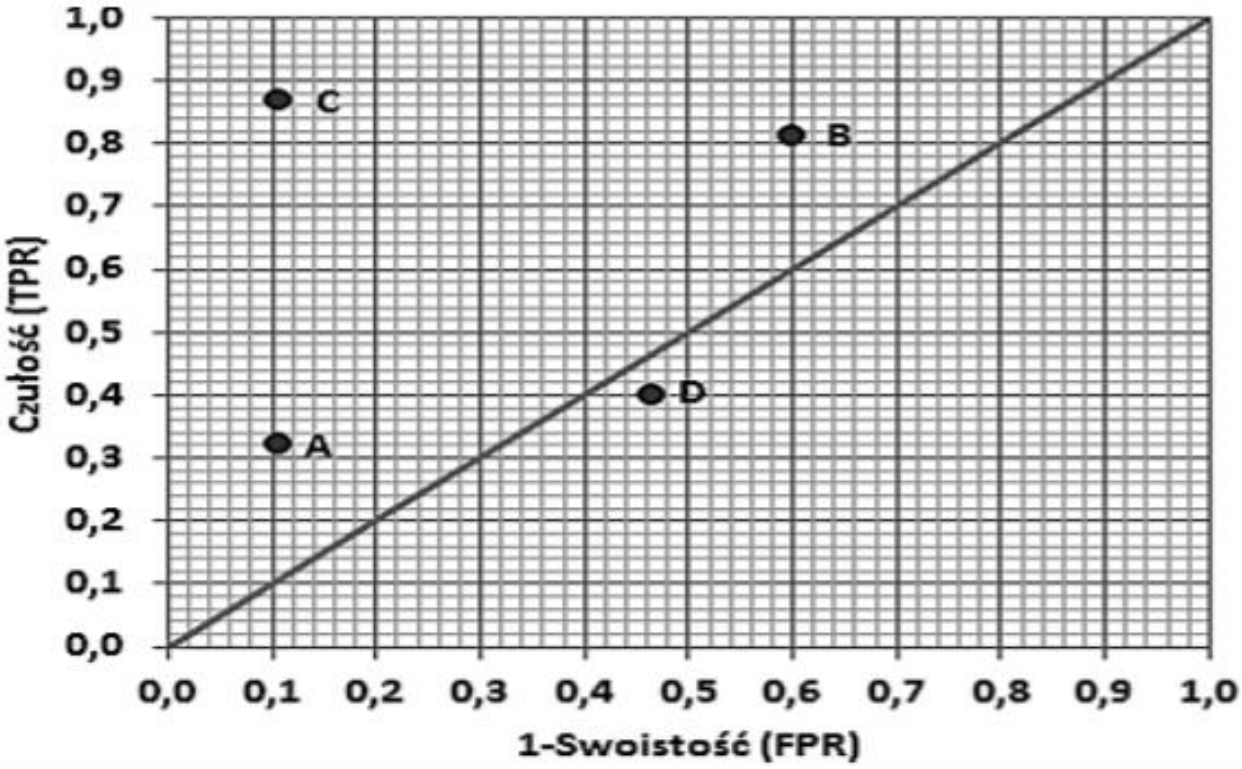
3.6 Krzywa ROC

W statystyce matematycznej „krzywa ROC jest graficzną reprezentacją efektywności modelu predykcyjnego poprzez wykreślenie charakterystyki jakościowej klasyfikatorów binarnych powstałych z modelu przy zastosowaniu wielu różnych punktów odcięcia”**.** Innymi słowy – każdy punkt krzywej ROC (*Receiver operating characteristic*) odpowiada innej macierzy błędu uzyskanej poprzez modyfikowanie „cut-off point” . Im więcej miejsc odcięcia zbadamy, tym więcej uzyskamy punktów na krzywej ROC. Zatem finalnie na wykres nanosimy *TPR* (czułość – oś pionowa) oraz *FPR* (1-specyficzność – oś pozioma)[156].

W przypadku klasyfikatorów dyskretnych opisywany jest pojedynczy punkt w przestrzeni ROC. Na rysunku 3.6 widoczne są przykładowe klasyfikatory dyskretne w przestrzeni ROC. Perfekcyjną klasyfikację przedstawia punkt o współrzędnych (0, 1). Punkty znajdujące się na przekątnej odpowiadają klasyfikacji losowej. Jeden punkt w przestrzeni krzywej ROC jest „lepszy” od drugiego, jeżeli jest zlokalizowany bardziej na „północny zachód” (C). Punkty umiejscowione bliżej lewej strony i bliżej osi X obrazują klasyfikatory bardziej konserwatywne (A), natomiast punkty położone bliżej górnej prawej strony opisują klasyfikatory bardziej liberalne (B). Z kolei punkt leżący poniżej przekątnej oznacza klasyfikację gorszą niż losowe przydzielanie obiektów do klas (D)[157].

[156] Blog matematyczny, https://mathspace.pl/matematyka/receiver-operating-characteristic-krzywa-roc-czyli-ocena-jakosci-klasyfikacji-czesc-7/ (dostęp: 26.02.2020).
[157] M. Misztal, *Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 2014, nr 328, s. 159.

Rys. 3.6. Porównanie czterech przykładowych klasyfikatorów dyskretnych w przestrzeni ROC



Źródło: M. Misztal, *Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 328, 2014, s. 159.

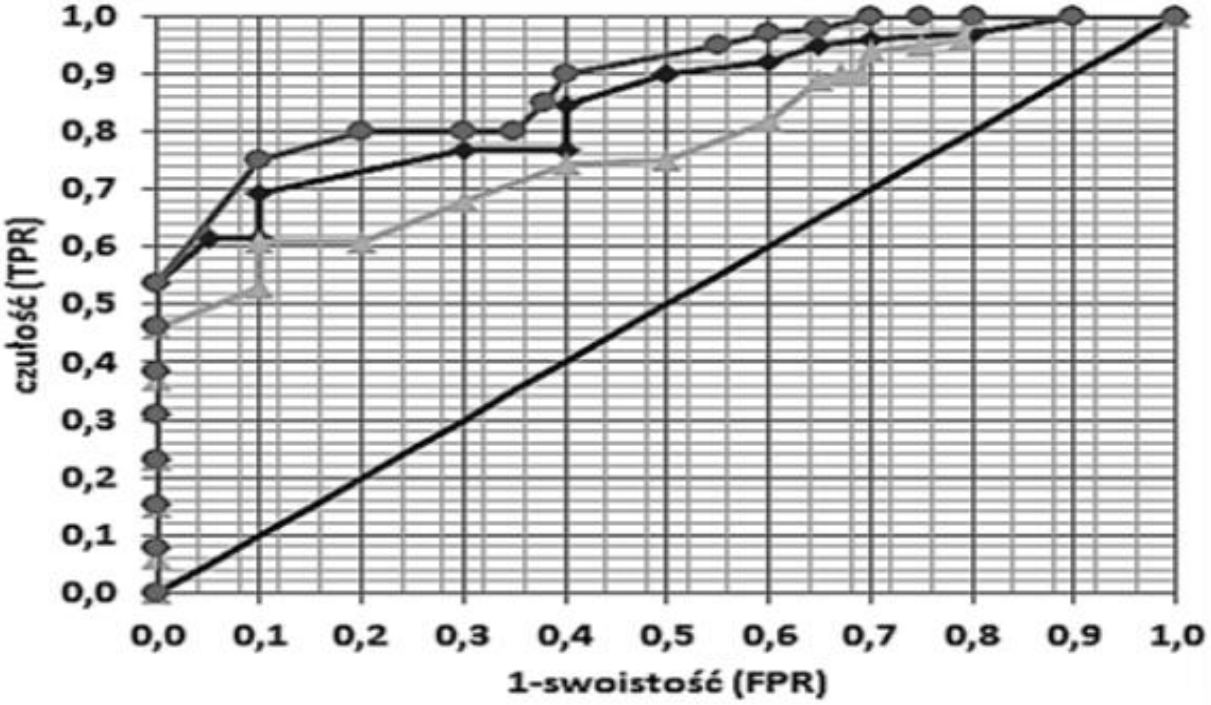
W przypadku klasyfikatorów ciągłych dla wszystkich punktów odcięcia należy obliczyć wartości FPR (1-swoistość) i TPR (czułość), umieścić je na wykresie, a potem połączyć, uzyskując krzywą ROC. Im wyżej położona krzywa ROC, tym lepsza zdolność predykcyjna modelu. Optymalny punkt odcięcia można określić na podstawie wskaźnika Youdena[158]:

$$Y = TPR - FPR. \qquad (3.18)$$

Na rysunku 3.7 zobrazowano przykładowe krzywe ROC dla klasyfikatorów ciągłych.

[158] Ibidem, s. 160.

Rys. 3.7. Porównanie trzech przykładowych klasyfikatorów ciągłych z wykorzystaniem krzywych ROC



Źródło: M. Misztal, *Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 328, 2014, s. 159.

Do oceny jakości klasyfikacji modelu na podstawie krzywej ROC można wyznaczyć pole pod wykresem krzywej – AUC (*area under the curve*) i traktować je jako miarę trafności i dobroci danego modelu. Jakość klasyfikacyjna modelu jest akceptowalna, gdy krzywa znajduje się powyżej przekątnej $y = x$, czyli gdy pole AUC jest większe od 0,5. W tym celu testuje się hipotezę zerową stanowiącą o tym, że pole pod wykresem krzywej ROC jest równe 0,5 (tj. wartości minimalnej). Wówczas statystyka testowa ma postać[159]:

$$Z = \frac{\widehat{AUC} - 0{,}5}{\sqrt{\hat{V}ar(\widehat{AUC})}}, \qquad (3.19)$$

gdzie $\hat{V}ar(\widehat{AUC})$ jest estymatorem wariancji pola $\widehat{AUC}$.

Statystyka Z ma asymptotycznie rozkład normalny (dla dużych liczności). Brak odrzucenia hipotezy zerowej stanowi o tym, że model nie ma żadnej mocy predykcyjnej.

[159] A. Sączewska-Piotrowska, *Zastosowanie krzywych ROC w analizie ubóstwa miejskich i wiejskich gospodarstw domowych*, „Przegląd Statystyczny", 2016, nr 2, s. 217.

Wartości AUC (*Area Under ROC Curve*) często są interpretowane następująco[160]:

- AUC ≥ 0,9 oznacza doskonały model;
- AUC ≥ 0,8 i < 0,9 model dobry;
- AUC ≥ 0,7 i < 0,8 akceptowalny model;
- AUC ≥ 0,6 i < 0,7 kiepski model;
- AUC ≥ 0,5 i < 0,6 model nieakceptowalny.

Pole AUC może być również wykorzystywane do porównania zdolności predykcyjnej różnych klasyfikatorów lub do oceny jakości tego samego klasyfikatora przed i po uwzględnieniu w modelu dodatkowych zmiennych.

[160] M. Szeliga, op. cit., s. 303.

Rozdział 4. Budowa modelu scoringowego – wyniki analizy empirycznej

Budowa modelu scoringowego jest jedną z najważniejszych i bardzo często najbardziej pracochłonną fazą projektowania całego systemu scoringowego. Faza ta obejmuje następujące czynności: wybór populacji bazowej (próby uczącej), określenie „dobrych" i „złych" klientów, analizę danych kredytowych oraz dobór właściwych predyktorów i odpowiednich ich atrybutów (grupowanie atrybutów, przekodowanie zmiennych). Etap ten obejmuje również wybór odpowiedniej metody estymacji modelu oraz przypisanie atrybutom predyktorów właściwych ocen punktowych, konstrukcję karty scoringowej oraz cały proces walidacji modelu (ocena jakości klasyfikacyjnej oszacowanego modelu, analiza właściwości prognostycznych).

Do budowy modeli scoringowych można zastosować zarówno metody statystyczne, jak i metody data mining. W ramach pierwszej grupy najczęściej wybieraną metodą jest regresja logistyczna. Z kolei najpopularniejszymi metodami eksploracji danych używanymi w scoringu są: sieci neuronowe, drzewa wzmacniane gradientowo (*gradient boosted trees*), losowy las (*random forest*) oraz metoda maszyn wektorów nośnych (*suport vector machines*).

Metody statystyczne charakteryzują się prostotą interpretacji, łatwością opisania siły i kierunku wpływu poszczególnych zmiennych na model. Bardzo ważną zaletą regresji logistycznej jest możliwość przekształcenia uzyskanego wzoru do formatu karty skoringowej, co ułatwia zrozumienie zbudowanego modelu nawet osobom niemających specjalistycznej wiedzy z obszaru analityki danych. Z kolei metody data mining działające na zasadzie „czarnej skrzynki", są dużo bardziej skomplikowane, przez co praktyczna interpretacja parametrów modelu jest niemożliwa. Odznaczają się jednak większą siłą predykcyjną, gdyż umożliwiają modelowanie nieliniowych wzorców bez potrzeby implementowania dodatkowych zmiennych reprezentujących zidentyfikowane interakcje[161].

4.1 Zdefiniowanie problemu i opis danych

W niniejszej pracy postanowiono porównać skuteczność 7 metod klasyfikacji służących do budowy modeli scoringowych: regresji logistycznej, drzew decyzyjnych, lasów losowych, drzew wzmacnianych gredientowo, maszyn wektorów wspierających, sieci neuronowych oraz algorytmu XGBoost. Do porównania jakości klasyfikacji zastosowanych

[161] J. Marcinkowska, *Metody statystyczne i eksploracji danych (data mining) w ocenie występowania omdleń w grupie częstoskurczu z wąskim zespołem QRS (AVNRT i AVRT*, Rozprawa doktorska, Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu, 2015, s. 52.

metod użyto krzywej ROC i miary AUC. Dodatkowo dla modelu regresji logistycznej z uwagi na pewne własności statystyczne pozwalające na interpretowanie ocen parametrów, zbudowano kartę scoringową oraz wyznaczono optymalny punkt odcięcia. Z kolei, dla najbardziej skutecznego z modeli data mining za pomocą wartości SHAP wyznaczono najbardziej istotne cechy kredytobiorcy oraz określono siłę i kierunek ich wpływu na zmienną prognozowaną (zdolność kredytową).

Modele zostały zbudowane w oparciu o dane klientów Home Credit Group[162]. Dane te zostały udostępnione społeczności www.kaggle.com przez Home Credit Group w ramach konkursu na budowę modelu prognozującego ryzyko braku spłaty udzielonej pożyczki.

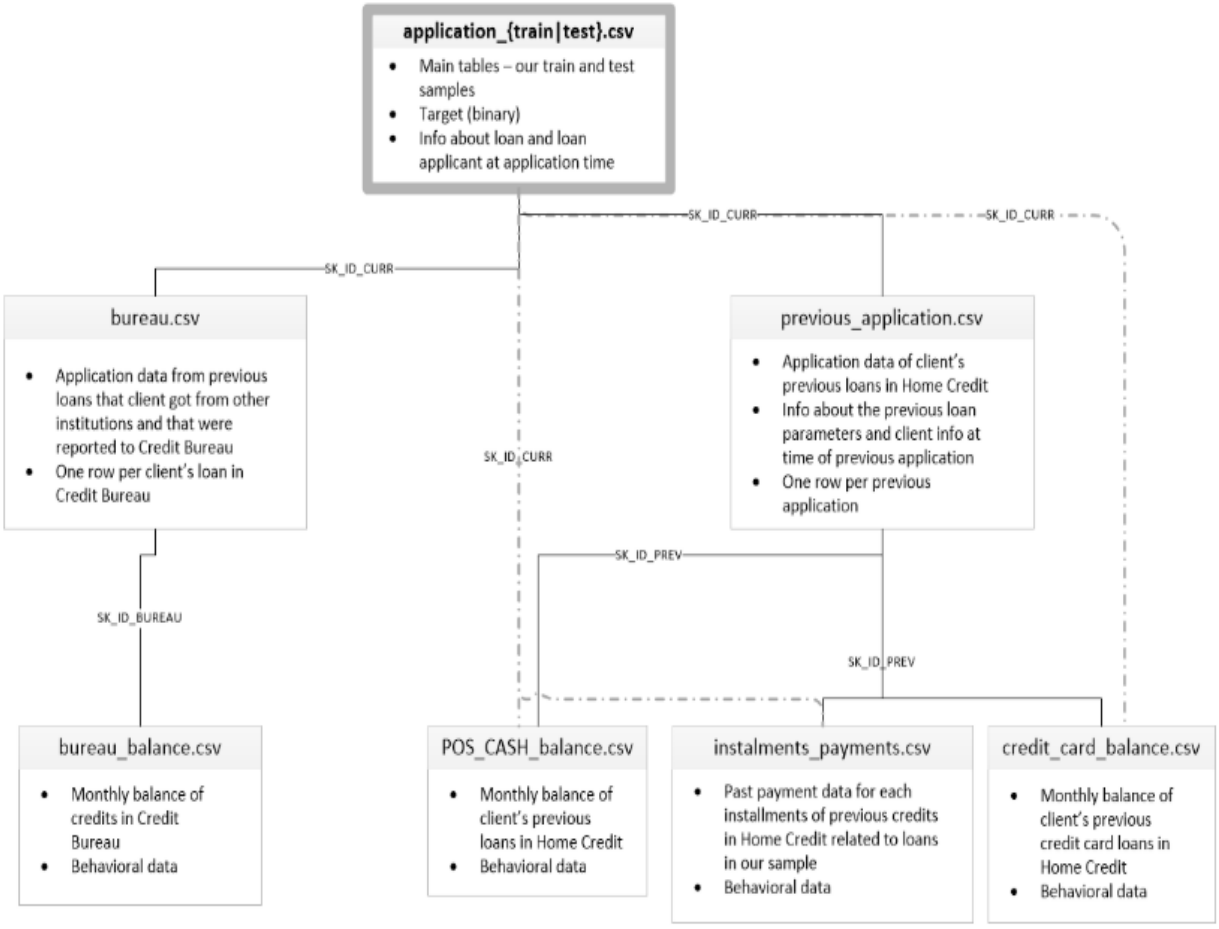
Zbiór danych podlegający analizie jest syntezą 7 baz danych, zawierających informacje pochodzące z:

- historii dotychczasowej współpracy z bankiem,
- informacji w bieżących i poprzednich wnioskach kredytowych,
- opinii i referencji z innych banków,
- biura informacji kredytowej.

Na rysunku 4.1 przedstawiono sieć połączeń jakich należy dokonać, aby uzyskać cały zbiór danych w formie jednej tabeli.

[162] Home Credit Group jest międzynarodową instytucją finansową zajmującą się głównie udzielaniem krótkoterminowych pożyczek konsumenckich.

Rys. 4.1. Schemat połączeń baz danych



Źródło: https://www.kaggle.com/c/home-credit-default-risk/data (dostęp: 18.05.2020).

Łącznie zbiór danych składa się z 307 tys. wierszy (kredytobiorców) oraz 214 kolumn (cech kredytobiorców). Ze zbioru danych postanowiono usunąć wszystkie kolumny, których braki wartości przekraczały 60%. W pozostałych kolumnach, gdzie występowały braki wartości, zostały one zastąpione w przypadku cech ilościowych średnią wartości odpowiednią dla każdej zmiennej, natomiast dla danych jakościowych dominantą. W wyniku oczyszczania danych usunięto 52 kolumny. W ostatecznej formie zbiór danych stanowi 307 tys. wierszy oraz 162 kolumny. Do dalszej analizy wykorzystano losową próbę 50 tys. obserwacji[163].

[163] Zbiór danych ograniczono do 50 tys. obserwacji ze względu na pewne ograniczenia techniczne przetwarzania danych w ramach dostępnego sprzętu komputerowego.

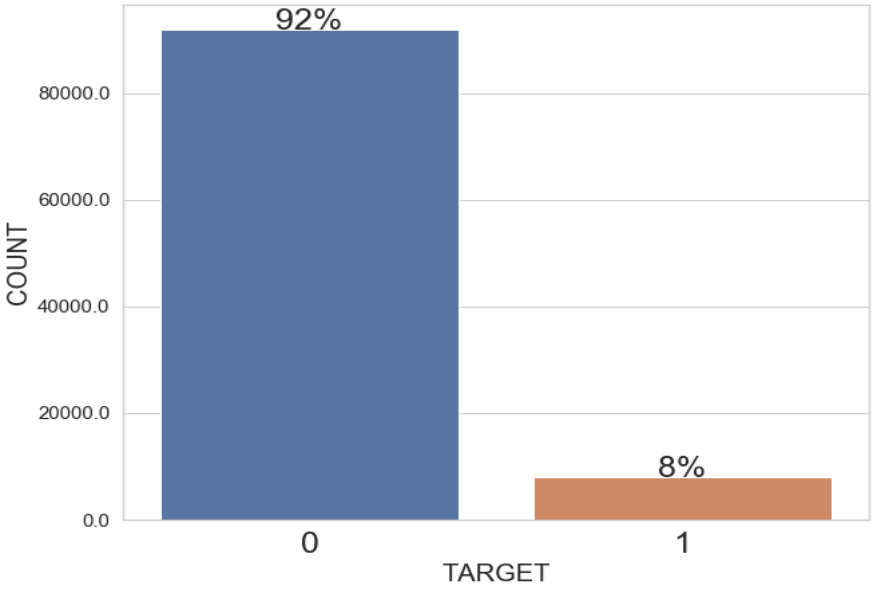
Tabela 4.1. Przykładowe zmienne z brakującymi wartościami

Zmienna	liczba zmiennych bez wartości	udział procentowy zmiennych bez wartości
STATUS_4	305484	99,30%
STATUS_5	305305	99,30%
STATUS_3	304766	99,10%
RATE_INTEREST_PRIVILEGED	302902	98,50%
RATE_INTEREST_PRIMARY	302902	98,50%
NUNIQUE_STATUS_y	264302	85,90%
AMT_PAYMENT_CURRENT	246451	80,10%
CNT_DRAWINGS_ATM_CURRENT	246371	80,10%
...	...	...

Źródło: opracowanie własne.

Klienta „złego", mającego problemy ze spłatą zaciągniętej pożyczki oznaczono za pomocą cyfry '1', natomiast klienta „dobrego" cyfrą '0'. Liczba klientów „dobrych" znacząco przewyższa liczbę klientów „złych", którzy stanowią tylko 8% wszystkich klientów.

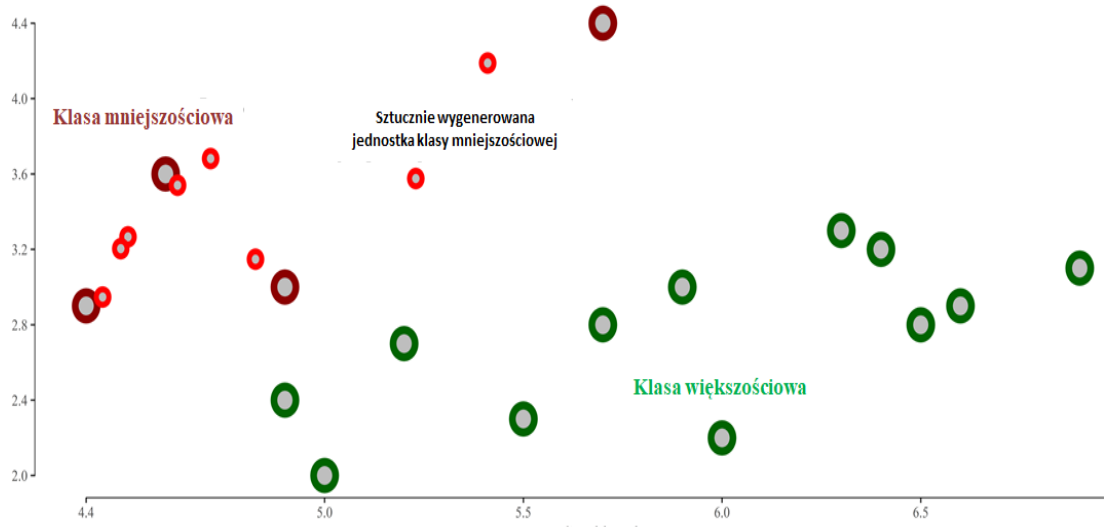
Wykres 4.1. Udział procentowy „dobrych" i „złych" klientów



Źródło: opracowanie własne.

W dalszym etapie pracy podjęta została próba rozwiązania problemu danych niezbilansowanych, poprzez zastosowanie jednego z algorytmów nadlosowujących (*oversampling*) obiekty klasy mniejszościowej – algorytmu SMOTE. Algorytm ten dogenerowuje nowe przykłady klasy mniejszościowej pomiędzy przykładami pierwotnymi za pomocą algorytmu k-najbliższych sąsiadów.

Rys. 4.2. Wizualizacja działania algorytmu SMOTE



Źródło: opracowanie własne.

4.2 Selekcja zmiennych i uczenie modeli

Doboru zmiennych do modelu dokonano metodą selekcji głosowania większościowego. Dla każdej zmiennej obliczono cztery współczynniki pozwalające ocenić istotność danej zmiennej: współczynnik wartości informacyjnej (IV), współczynnik redukcji zanieczyszczenia lasów losowych (*Random Forest*), współczynnik rekurencyjnej eliminacji wstecznej (REF) oraz współczynnik Chi^2. W przypadku, gdy co najmniej 3 współczynniki okazały się istotne, zmienną dodano do modelu.

W tabeli 4.2 przedstawiono ranking wstępnie wytypowanych predyktorów, które zostaną wykorzystane do konstrukcji modelu scoringowego, uporządkowanych ze względu na wartości współczynnika IV. Analizując wartości wskaźników można zauważyć, że predyktory wykazują średnie zdolności klasyfikacyjne. Przyjmuje się, że wartości IV powyżej 0,3 wskazują na silną moc predykcyjną, natomiast wartości poniżej 0,02 na całkowity brak mocy predykcyjnej[164]. Do dalszej analizy wybrano zatem 54 zmienne, których własności klasyfikacyjne zostały potwierdzone przez istotność 3 współczynników. W tabeli 4.3 ukazano przykładowe zmienne wchodzące w skład modelu, uszeregowane ze względu na liczbę istotnych współczynników.

[164] Cioch K., Karnowska K., *Ocena modeli scoringowych w SKOK Stefczyka*, StatSoft, Kraków, 2010, http://www.statsoft.pl/czytelnia/artykuly/Ocena_modeli_skoringowych_w_SK OK_Stefczyka. pdf (dostęp: 23.05.2020).

Tabela 4.2. Wartości współczynników metod selekcji zmiennych

Nazwa zmiennej	IV	Random_Forest	REF	Chi_Square
EXT_SOURCE_3	0,265	0,044	-2,717	60,965
EXT_SOURCE_2	0,239	0,039	-2,103	79,139
DAYS_CREDIT	0,132	0,017	0	9,688
DAYS_EMPLOYED	0,128	0,02	0	4,145
EXT_SOURCE_1	0,119	0,02	-1,388	15,38
AGE_CLIENT	0,097	0,021	0	21,007
DAYS_CREDIT_ENDDATE	0,092	0,016	0	0,532
NAME_INCOME_TYPE	0,088	0,005	0,417	25,373
NAME_EDUCATION_TYPE	0,088	0,005	0,513	16,843
…	…	…	…	…

Źródło: opracowanie własne.

Tabela 4.3. Oceny istotności współczynników w formie „głosu”

Nazwa zmiennej	IV	Random_Forest	REF	Chi_Square	Głosy
EXT_SOURCE_3	1	1	1	1	4
NAME_INCOME_TYPE	1	1	1	1	4
REGION_POPULATION_RELATIVE	1	1	1	1	4
OCCUPATION_TYPE	1	1	1	1	4
NAME_EDUCATION_TYPE	1	1	1	1	4
EXT_SOURCE_2	1	1	1	1	4
EXT_SOURCE_1	1	1	1	1	4
NUM_INSTALMENT_VERSION	1	0	1	1	3
AMT_ANNUITY_y	1	1	0	1	3
…	…	…	…	…	…

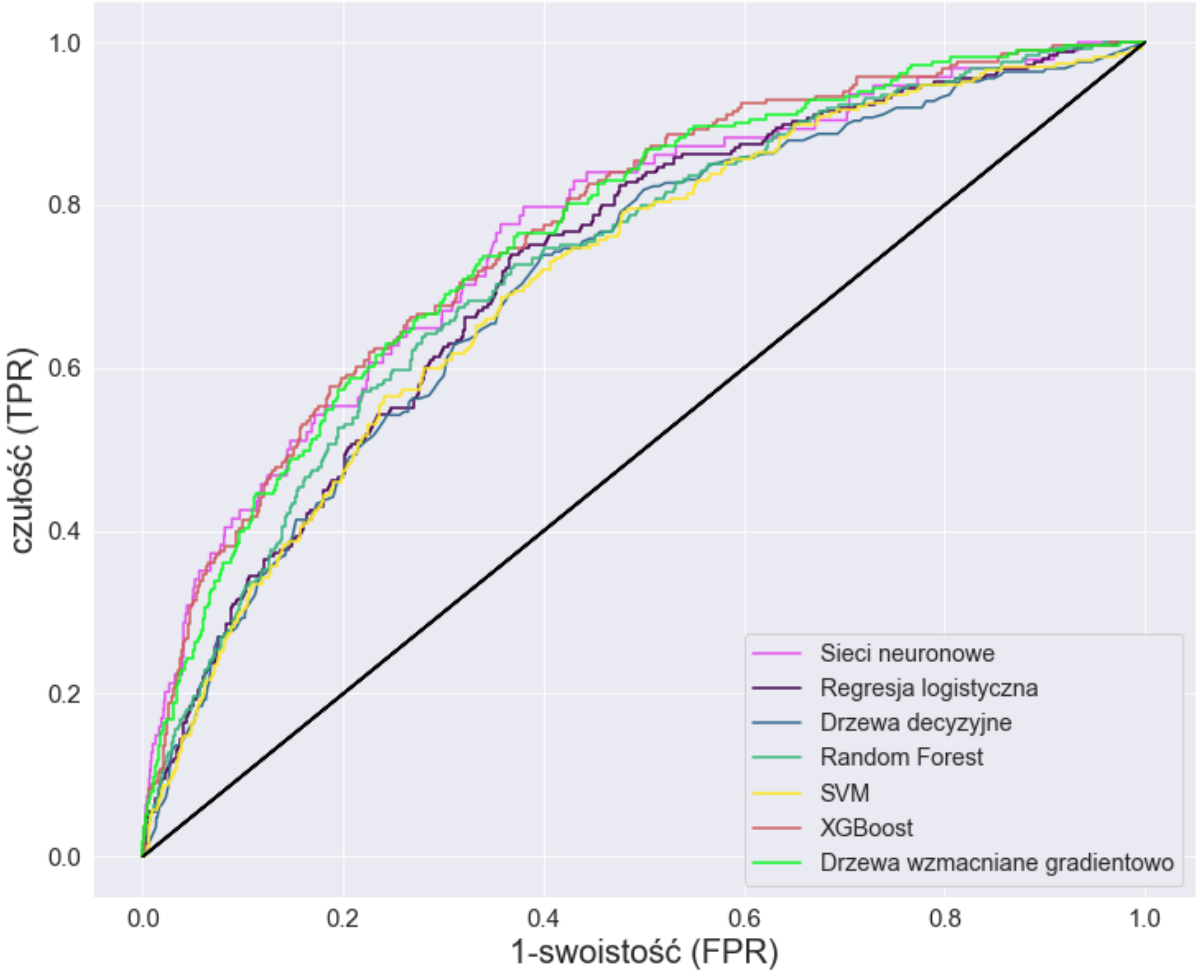
Źródło: opracowanie własne.

Przed przystąpieniem do modelowania zbiór danych podzielono na zbiór treningowy oraz zbiór testowy w proporcji 80:20. Zbiór treningowy poddano przekształceniu zgodnie z algorytmem SMOTE w celu uzyskania zbilansowanej próby. Następnie na zbiorze treningowym dokonano budowy 7 modeli scoringowych w oparciu o wcześniej opisane metody. Proces uczenia każdego modelu przeprowadzono za pomocą 5-krotnego sprawdzianu krzyżowego, wydzielając dodatkowy zbiór walidacyjny do testowania skuteczności modelu. Jednocześnie zastosowano metodę przeszukiwania siatki w celu znalezienia optymalnych hiperparametrów stosowanego algorytmu klasyfikacji. Po odpowiednim wytrenowaniu modelu dokonano prognozowania na wcześniej wyodrębnionej niezależnej próbie testowej, zachowującej pierwotny rozkład zmiennej prognozowanej.

4.3 Porównanie i ocena modeli

Otrzymane wyniki analiz, dokonanych przy użyciu różnych metod, porównano za pomocą krzywej ROC. Na podstawie przebiegu krzywej ROC oraz wyliczonych miar AUC stwierdzono, że wszystkim modelom udało się wychwycić zależności zawarte w danych – pole powierzchni dla krzywej ROC wynosi powyżej 0,7. Modele mają porównywalną jakość klasyfikacji, jednak największą skuteczność klasyfikacji uzyskuje model XGBoost, którego wartość AUC wynosi 0,77. Najmniej skuteczny okazał się model drzew decyzyjnych (AUC – 0,71). Przebieg krzywej ROC dla każdego z modeli widoczny jest na wykresie 4.2, natomiast wszystkie wyniki miary AUC zawarte są w tabeli 4.4.

Wykres 4.2. Krzywe ROC dla analizowanych modeli



Źródło: opracowanie własne.

Tabela 4.4. Wartości AUC dla analizowanych modeli

Nazwa modelu	AUC
XGBoost	0,771
Drzewa wzmacniane gradientowo	0,765
Sieci neuronowe	0,764
Random Forest	0,73
Regresja logistyczna	0,726
SVM	0,714
Drzewa decyzyjne	0,711

Źródło: opracowanie własne.

Wiarygodność i trafność predykcji są kluczowe w procesie walidacji modelu, ale nie powinny być jedynymi kryteriami oceny modeli uczenia maszynowego. Coraz częściej dużą wagę przywiązuje się do interpretowalności predykcji. Po pierwsze dlatego, że traktowanie modeli data mining jak czarnych skrzynek lub magicznych wyroczni ogranicza zaufanie użytkowników do skuteczności tych metod. Po drugie, wraz z coraz powszechniejszym wykorzystaniem modeli uczenia maszynowego wzrasta liczba ich zastosowań, w których wyjaśnienie podjętej decyzji jest równie ważne co sama decyzja. Na przykład, bank ma prawny obowiązek podać powód odmowy udzielenia kredytu.

Znalezienie uniwersalnej, znajdującej zastosowanie dla dowolnego modelu uczenia maszynowego, metody tłumaczącej prawidłowości predykcji jest przedmiotem intensywnych, trwających od ponad 20 lat badań. Efektem tych prac jest opracowanie dwóch głównych metod wyjaśniania[165]:

- metoda LIME (*Local Interpretable Model-Agnostic Explanations*) – ukryty model jest aproksymowany przy pomocy modelu czytelnego, który podejmując decyzję bazuje na modyfikacjach oryginalnego wejścia. Predykcje prostszego, nauczonego na kilku, góra kilkudziesięciu przykładach, modelu można odpowiednio wyjaśnić. W tym celu porównuje się wyniki predykcji lokalnego modelu dla zmodyfikowanego na wiele sposobów przykładu;
- metoda SHAP (*Shapley Additive Explanation*) – istota tej metody polega na obliczeniu wpływu wartości poszczególnych zmiennych na wynik określonej predykcji. Stosowanych jest kilka technik dekompozycji wpływu poszczególnych zmiennych na uzyskaną prognozę. Najprostsza polega na zastąpieniu wartości określonej zmiennej

[165] Miesięcznik dla specjalistów z branży IT, http://www.it-professional.pl/stacje-robocze/artykul,8536,automatyczne-uczenie-maszynowe-z-usluga-azure-ml.html (dostęp: 26.05.2020).

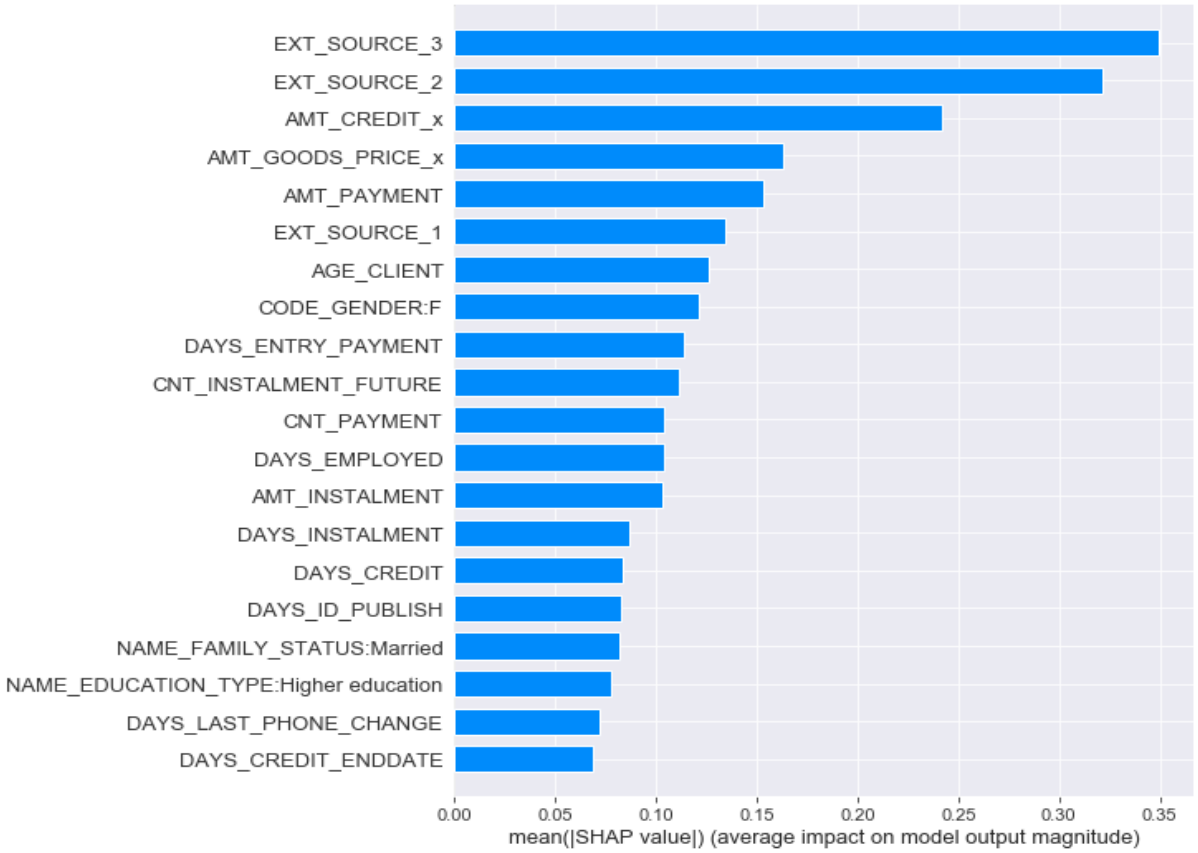
wartością losową, wielokrotnym powtórzeniu predykcji i porównaniu otrzymanych wyników.

4.3.1 Analiza modelu scoringowego - XGBoost

Dla modelu o najlepszych własnościach klasyfikacyjnych (XGBoost) postanowiono rozszerzyć analizę o interpretację uzyskanych wyników. Za pomocą metody SHAP wyznaczono najbardziej istotne zmienne oraz określono siłę i kierunek ich oddziaływania na zmienną prognozowaną.

Sumaryczne wartości SHAP mogą pokazać, w jakim stopniu każda cecha przyczyniła się do przewidywania zmiennej docelowej. Poniżej zaprezentowano wykres z globalną interpretowalnością (wykres 4.3). Wykres ten zawiera listę najbardziej znaczących zmiennych w porządku malejącym. Górne charakterystyki wnoszą najwięcej do modelu. Im niżej tym cecha ma mniejszą moc predykcyjną.

Wykres 4.3. Ocena mocy dyskryminacyjnej zmiennych według wartości SHAP

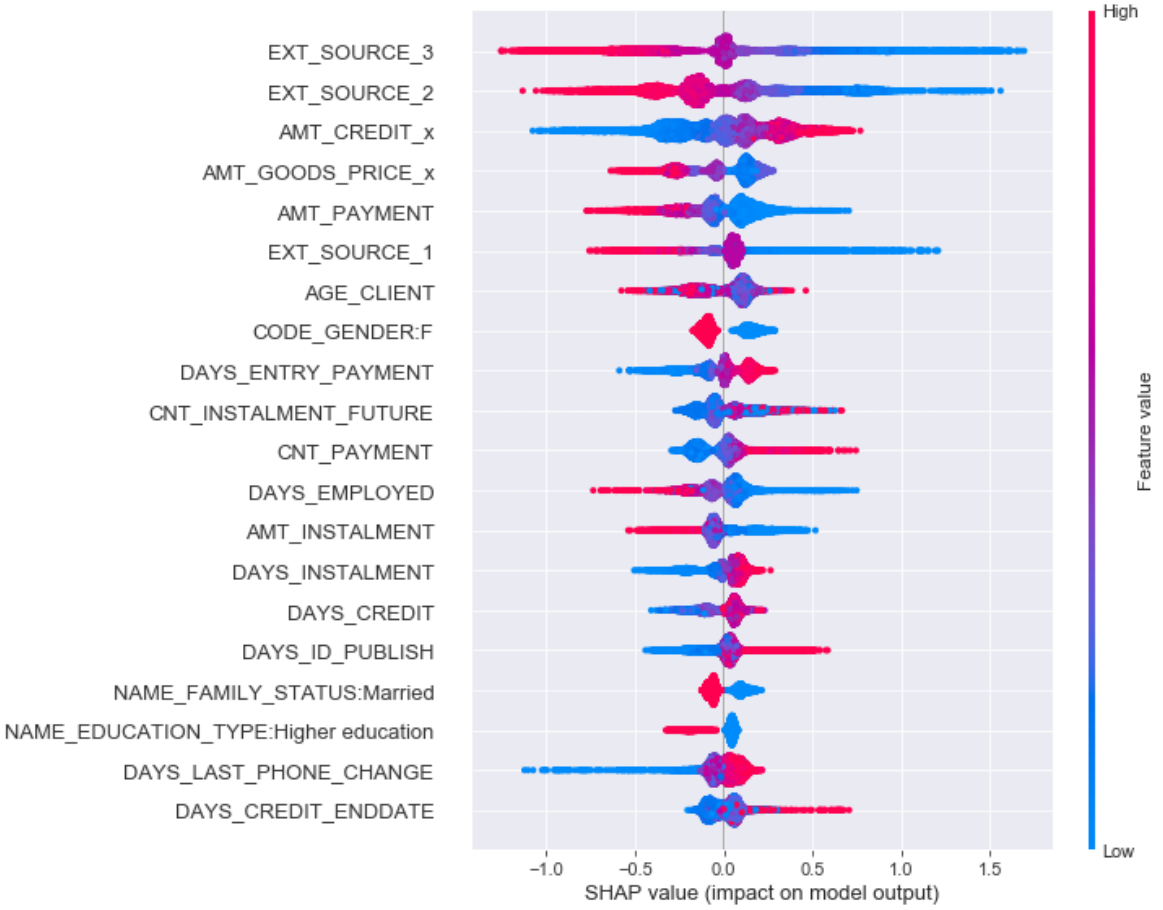


Źródło: opracowanie własne.

Kolejny wykres wartości SHAP (wykres 4.4) ma na celu pokazanie kierunku i siły oddziaływania poszczególnych zmiennych na zmienną prognozowaną (zdolność kredytową).

Zachowana jest ta sama kolejność zmiennych według ważności cech. Każda kropka na wykresie reprezentuje jedną obserwację w zbiorze danych (w tym wypadku zbiór testowy).

Wykres 4.4. Globalne odziaływanie zmiennych na wynik predykcji



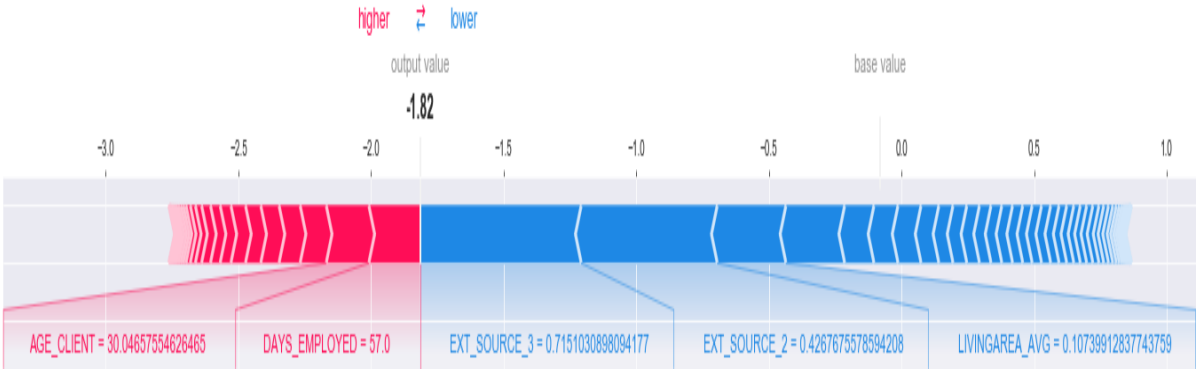
Źródło: opracowanie własne.

Oś pozioma na dole wykresu wskazuje, czy wpływ wartości SHAP określonych zmiennych jest związany z wyższą prognozą czy niższą (wszystko na lewo od 0 wpływa negatywnie, a wszystko na prawo pozytywnie). Indeks koloru widoczny po prawej stronie wykresu informuje jakie wartości przyjmuje dana zmienna (wysokie wartości kolor czerwony, niskie wartości kolor niebieski). Dodatkowo poprzez układ kropek możemy odczytać uproszczony rozkład określonej zmiennej. Znaczne zagęszczenie kropek w jednym miejscu oznacza, że większość obserwacji jest skupiona w tym jednym miejscu.

Wykres można interpretować w następujący sposób – pozytywnie na zdolność kredytową wpływa posiadanie wyższego wykształcenia oraz bycie w związku małżeńskim, a negatywnie krótki okres zatrudnienia i duża wielkość zaciągniętej pożyczki.

Podczas przeliczania każda obserwacja otrzymała swój własny zestaw wartości SHAP. Dzięki temu dla każdej obserwacji możemy sprawdzić w jakim stopniu każda zmienna wpłynęła na podjęcie takiej decyzji przez model. Możemy w prosty i przejrzysty sposób wyjaśnić, dlaczego model zwrócił taką prognozę. Poniżej przedstawiono wykres SHAP dla pierwszej obserwacji ze zbioru testowego (wykres 4.5).

Wykres 4.5. Lokalne oddziaływanie zmiennych na wynik predykcji



Źródło: opracowanie własne.

Dla analizowanej obserwacji (klienta) wartość prognozowana (*output value*) wynosi -1.82 i jest ona zapisana w postaci log-odds. Aby przejść do formatu prawdopodobieństwa należy dokonać przekształcenia za pomocą wzoru:

$$probability = \frac{\exp(\alpha)}{1 + \exp(\alpha)}. \qquad (4.1)$$

Wówczas prawdopodobieństwo braku spłaty zaciągniętego kredytu dla tego klienta wynosi 14%.

Zmienne (cechy) podnoszące prognozę (zwiększające prawdopodobieństwo niespłacenia kredytu) są pokazane na czerwono, a te które zmniejszają prognozę na niebiesko. Cechy są poustawiane w kolejności od najbardziej oddziałujących odpowiednio na plus i minus.

Największy wpływ na otrzymany wynik prognozy, zwiększający prawdopodobieństwo braku spłaty kredytu miały liczba dni zatrudnienia klienta w obecnej pracy (57 dni) oraz jego wiek (30 lat). Z kolei najbardziej istotnymi zmiennymi, które zdecydowały o niskim poziomie prawdopodobieństwa były znormalizowane wartości

świadczące o warunkach zamieszkania klienta oraz dwie znormalizowane zmienne pochodzące z zewnętrznych baz danych[166].

4.3.2 Analiza modelu scoringowego – regresja logistyczna

Metoda regresji logistycznej pozwala na przedstawienie modelowanych zależności w postaci ilorazu szans, co umożliwia przekształcenie uzyskanego wyniku do formatu punktacji scoringowej i zbudowanie karty scoringowej.

Ponieważ tablica scoringowa jest narzędziem stosowanym dla danych dyskretnych (każda z cech uwzględnianych w modelu jest podzielona na przedziały), pierwszym etapem procesu przygotowania tablicy przed przystąpieniem do modelowania była odpowiednia modyfikacja danych, czyli dyskretyzacja zmiennych ciągłych i kategoryzacja zmiennych jakościowych. Proces dyskretyzacji przeprowadzono poprzez przekształcenia interakcyjne, bazując na analizie stosowanych raportów informujących o wartościach współczynników WOE (*Weight of Evidence*) i IV (*Information Value*) oraz na podstawie wykresu WOE.

Tabela 4.5. Przykładowe wartości współczynników WOE

Nazwa zmiennej	Kategoria	WOE
…	…	…
AMT_DOWN_PAYMENT	(7068.265, 832500.0]	0,28413
AMT_DOWN_PAYMENT	(-0.001, 1771.2]	-0,18368
AMT_DOWN_PAYMENT	(1771.2, 7068.265]	0,002382
CODE_GENDER	F	0,134396
CODE_GENDER	M	-0,22332
NAME_INCOME_TYPE	Commercial associate	0,146697
NAME_INCOME_TYPE	Pensioner	0,417076
NAME_INCOME_TYPE	Unemployed	-1,0455
…	…	…

Źródło: opracowanie własne.

Po zbudowaniu modelu regresji logistycznej kolejnym etapem budowy karty scoringowej było określenie minimalnej i maksymalnej wartości granicznej punktacji scoringowej. W tym wypadku zdecydowano się przyjąć jako wartość minimalną 300 punktów, natomiast dla wartości maksymalnej 850 punktów. Następnie według wzoru 4.2 w

[166] W opisie zbioru danych dostarczonym przez Home Credit Group nie znajdujemy szczegółowych informacji na temat zmiennych: EXT_SOURCE_2 i EXT_SOURCE_3. Natomiast na podstawie dużej zależności tych zmiennych wobec zmiennej prognozowanej możemy podejrzewać, że przedstawiają one jakiś znormalizowany współczynnik ryzyka kredytowego szacowany w ramach poprzednich pożyczek.

oparciu o uzyskany współczynnik regresji obliczona została punktacja scoringowa dla każdej zdyskretyzowanej zmiennej[167]:

$$variable\_score = variable\_coef \frac{(\max\_score - \min\_score)}{(\max\_sum\_coef\ - \min\_sum\_coef)} \quad . \tag{4.2}$$

W dalszej kolejności dokonano normalizacji wartości wyrazu wolnego (*intercept*) przy pomocy wzoru[168]:

$$intercept\_score = \frac{(intercept\_coef - \min\_score)}{(\max\_sum\_coef - \min\_sum\_coef)} (\max\_score - \min\_score) + \min\_score \, . \tag{4.3}$$

Suma punktów scoringowych dla określonych cech kredytobiorcy stanowi wynik scoringowy, od którego zależy, czy bank zdecyduje się na podjęcie decyzji o udzieleniu kredytu. Fragment karty scoringowej przedstawiono w tabeli 4.6.

Tabela 4.6. Fragment karty scoringowej

Nazwa zdyskretyzowanej zmiennej	Współczynnik regresji	Oryginalna nazwa zmiennej	Punktacja
…	….	….	….
AMT_CREDIT_x:(44999.999, 337500.0]	0,276	AMT_CREDIT_x	10
AMT_CREDIT_x:(337500.0, 679500.0]	0,02	AMT_CREDIT_x	1
AMT_CREDIT_x:(679500.0, 4050000.0]	-0,185	AMT_CREDIT_x	-7
OCCUPATION_TYPE:Cleaning staff	-0,06	OCCUPATION_TYPE	-2
OCCUPATION_TYPE:Cooking staff	0,025	OCCUPATION_TYPE	1
OCCUPATION_TYPE:Core staff	0,16	OCCUPATION_TYPE	6
OCCUPATION_TYPE:Drivers	-0,107	OCCUPATION_TYPE	-4
OCCUPATION_TYPE:HR staff	0,166	OCCUPATION_TYPE	6
OCCUPATION_TYPE:High skill tech Staff	0,219	OCCUPATION_TYPE	8
….	….	….	….

Źródło: opracowanie własne.

Ostatnim elementem związanym z budową karty scoringowej jest wybór optymalnego punktu odcięcia. Punkt ten określa wartość graniczną punktacji skoringowej, poniżej której kredytobiorcę uznaje się za niezdolnego do spłaty kredytu (kredytobiorca zły). Istotnym problemem jest ustalenie tzw. optymalnego punktu „cut–off", który dzieli kredytobiorców na dwie klasy: „złych" o wysokim ryzyku niespłacenia kredytu oraz „dobrych" o niskim ryzyku.

[167] Kurs modelowania ryzyka kredytowego w języku Python, https://365datascience.com/courses/credit-risk-modeling-in-python/ (dostęp: 29.05.2020).
[168] Ibidem.

Przyjęcie zbyt niskiego scoringu jako punktu odcięcia powoduje z jednej strony wysoki wskaźnik tzw. stopy akceptacji (ogólnego odsetka przyznanych pożyczek). Z drugiej strony, przyczynia się do tego, że możemy mieć do czynienia z niską efektywnością II rodzaju, czyli niskim poziomem właściwego rozpoznawania złych kredytobiorców. Wysoki będzie także wtedy poziom złych kredytów w danej klasie (co jest niekorzystne szczególnie dla klasy dobrych kredytobiorców). Wyznaczenie asekuracyjnie dużej wartości scoringu jako punktu odcięcia skutkuje, że zwiększa się efektywność II rodzaju (co jest korzystne, gdyż zwiększa się liczba poprawnie zakwalifikowanych złych kredytów), ale tym samym zmniejsza się efektywność I rodzaju oraz maleje stopa akceptacji (mniejsza liczba klientów dla banku)[169].

Tabela 4.7 obrazuje efektywność klasyfikacyjną zbudowanego modelu scoringowego dla różnych wyznaczonych punktów odcięcia. Wyraźnie widać, jak przyjęty punkt odcięcia ma wpływ na poziom wniosków zaakceptowanych oraz odrzuconych, czy też na efektywność rozróżniania kredytobiorców spłacających i niespłacających kredyty. Widoczna jest dodatnia korelacja między prawdopodobieństwem niespłacenia kredytu a stopą akceptacji kredytobiorcy (wraz ze wzrostem ryzyka kredytowego zwiększa się poziom akceptacji). Na przykład, w sytuacji zastosowania punktu odcięcia na poziomie punktacji 638 poziom zaakceptowanych wniosków kredytowych wynosi 58,6%, a prawdopodobieństwo niespłacenia kredytu wynosi 10%. Z kolei, przy punktacji 542 stopa akceptacji wynosi 82%, natomiast prawdopodobieństwo braku spłaty pożyczki wynosi 20%.

Tabela 4.7. Efektywność klasyfikacji modelu scoringowego

Score	Stopa akceptacji	Stopa odrzucenia	Prawdopodobieństwo niespłacenia kredytu
…	…	…	…
702	**23,20%**	**76,80%**	**5%**
638	**58,60%**	**41,50%**	**10%**
593	**71,10%**	**29,00%**	**15%**
542	**82%**	**17,70%**	**20%**
…	…	…	…

Źródło: opracowanie własne.

Pożyczkodawca sam musi określić, na jakim poziomie ustali punkt progowy, uwzględniając przy tym poziom ryzyka, które jest w stanie zaakceptować. Określenie

[169] T. Pisula, *Metodyczne aspekty zastosowania modeli skoringowych do oceny zdolności kredytowej z wykorzystaniem metod ilościowych,* Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 2013, nr 33, s. 281-283.

wartości punktu odcięcia zależy w głównej mierze od prowadzonej polityki kredytowej oraz jakości aktywów banku.

Po wdrożeniu modelu do systemu ostatnią fazą praktycznego wykorzystania modeli scoringowych w procesie zarządzania ryzykiem kredytowym jest faza monitorowania. Głównym zadaniem monitoringu modeli skoringowych jest kontrolowanie, czy model nie potrzebuje korekty. Model będzie wymagał aktualizacji, gdy populacja nowych wnioskodawców zacznie się szybko zmieniać i będzie istotnie odbiegać od populacji bazowej, w oparciu o którą projektowano i wdrażano model. Jednym z istotniejszych etapów w zakresie fazy monitoringu modeli scoringowych jest badanie stabilności populacji oraz analiza zmian w rozkładzie cech charakteryzujących kredytobiorców, które użyto w modelu jako predyktory[170]. W niniejszej pracy ze względu na fakt, że stabilność populacji i cech kredytobiorcy bada się w dłuższym przedziale czasowym i wynikający z tego brak dostępnych danych, analiza monitoringu modelu scoringowego nie została przeprowadzona.

[170] Ibidem, s. 284-285.

Zakończenie

Zastosowanie modeli scoringowych pozwala bankom ograniczyć ryzyko kredytowe i przynosi im wymierne korzyści. Poprawia jakość portfela kredytowego lub przynajmniej pozwala zahamować negatywne tendencje w polityce kredytowej. Modele scoringowe wpływają na szybkość, bezpieczeństwo i skuteczność podejmowanych decyzji kredytowych. Ponadto umożliwiają kontrolę nad procesem kredytowania i przewidywanie złych długów.

Zbudowanie dobrego systemu scoringowego nie jest zadaniem prostym i wymaga zwykle bardzo przemyślanego zaprojektowania całego przedsięwzięcia. Efektywność i skuteczność w dużej mierze zależy od bardzo wielu różnych czynników. Pierwszym znaczącym czynnikiem warunkującym jego skuteczność jest dobór odpowiedniej próby badawczej do estymowanych modeli oraz wybór właściwych predyktorów mających dobre własności prognostyczne w klasyfikacji „dobrych" i „złych" kredytobiorców. Efektywność modeli scoringowych determinowana jest także wyborem odpowiedniej metody uczenia oraz właściwym strojeniem modelu, zapobiegającym nadmiernemu dopasowaniu. Skuteczność klasyfikacji modeli scoringowych oszacowanych różnymi metodami może niekiedy znacznie się różnić. W pracy wykorzystano 7 metod klasyfikacji do budowy modeli scoringowych. Do porównania skuteczności zastosowanych metod użyto krzywej ROC i miary AUC. Zgodnie z postawioną hipotezą, klasyczne metody statystyczne okazały się posiadać gorsze własności klasyfikacji niż metody data mining. Skuteczność klasyfikacji regresji logistycznej określona miarą AUC wyniosła 0,726. Słabszy wynik klasyfikacji uzyskały jedynie algorytmy maszyn wektorów nośnych (0,714) i drzewa decyzyjne (0,711). Z kolei najbardziej skutecznymi klasyfikatorami okazały się metody ze wzmacnianiem gradientowym, odpowiednio algorytm XGBoost o wartości AUC 0,771 i drzewa wzmacniane gradientowo, dla których pole pod powierzchnią krzywej ROC wyniosło 0,765.

Modele data mining są bardziej złożone i pozwalają szacować nieliniowe i często ulotne wzorce, co zwykle przekłada się na budowę lepszych modeli o większej sile predykcyjnej. Argumentem przemawiającym za stosowaniem metod statystycznych w budowaniu modeli scoringowych jest ich łatwość interpretacji. Własności statystyczne regresji logistycznej pozwalają na budowę kraty scoringowej na podstawie, której możliwe jest wypracowanie ustandaryzowanego postępowania w procesie udzielania kredytu. Metody eksploracji danych nie posiadają zdolności do bezpośredniej interpretacji parametrów modelu, natomiast możliwe jest zmierzenie siły i kierunku wpływu analizowanych zmiennych na

decyzję klasyfikacji, zarówno w wymiarze globalnym jak i lokalnym. Na podstawie modelu XGBoost w oparciu o metodę SHAP wyznaczone zostały cechy kredytobiorcy mające największy wpływ na ocenę zdolności kredytowej. Do najbardziej dyskryminacyjnych zmiennych zaliczono m.in. dwie znormalizowane zmienne pochodzące z zewnętrznych baz danych, wielkość zaciągniętego kredytu, cenę dobra na które zaciągnięto pożyczkę, kwotę spłaty poprzedniego kredytu oraz wiek klienta.

Dzięki przeprowadzeniu poprawnej oceny kredytobiorcy, jego zwyczajów płatniczych w zależności od bieżącej sytuacji finansowej, banki są w stanie ograniczyć ponoszone ryzyko. W sposób świadomy mogą określić, jaki poziom ryzyka będzie korzystny dla prowadzonej polityki kredytowej banku. Nie należy zapominać, że podstawowym celem działalności bankowej jest maksymalizacja zysku.

Literatura

1. Baster P., Pocztowska K., *Sieci neuronowe i polichotomiczne modele zmiennych jakościowych w analizie ryzyka kredytowego*, „Folia Oeconimca Cracoviensia”, 2011, vol. LII.
2. Branicka Z., *Metody konstrukcji oraz symulacyjne badanie właściwości jednorodnych i niejednorodnych komitetów klasyfikatorów,* Praca magisterska, Uniwersytet Warszawski Wydział Matematyki, Informatyki i Mechaniki, 2001.
3. Breiman L., Friedman J.H., Olshen R.A., Stone C.J., *Classification and Regression Trees*. Wadsworth International Group, Monterey,Ca, 1984.
4. Breiman L., *Random Forests*, „Machine Learning”, 2001, nr 45.
5. Breiman L. i in., *Classification and Regression Trees,* Wadsworth International Group, Monterey,Ca, 1984.
6. Bujak Ł., *Drzewa decyzyjne,* http://www.is.umk.pl/~duch/Wyklady/CIS/Prace%20zalicz/08-Bujak.pdf.
7. Burda A., *Prognozowanie kondycji ekonomiczno-finansowej przedsiębiorstw z wykorzystaniem sztucznych sieci neuronowych,* „Barometr Regionalny”, 2006, nr 6.
8. Chen T., Guestrin C., *XGBoost: A Scalable Tree Boosting System,* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, ACM Press, New York, 2016.
9. Cioch K., Karnowska K., Ocena modeli scoringowych w SKOK Stefczyka, StatSoft, Kraków, 2010, http://www.statsoft.pl/czytelnia/artykuly/Ocena_modeli_skoringowych_w_SKOK_Stefczyka. pdf.
10. Danieluk B., *Zastosowanie regresji logistycznej w badaniach eksperymentalnych,* Psychologia Społeczna, 2010, tom 5, nr 2-3 (14).
11. Darłak B., Włodarczyk M., *Zastosowanie sztucznej sieci neuronowej do uzupełnienia danych zbiornikowych*, „Przegląd Geologiczny”, 2001, vol. 49, nr 9.
12. Dawidowicz J., *Badania struktur sieci neuronowych typu MLP do oceny układu stref ciśnienia systemu dystrybucji wody*, „Civil and Environmental Engineering/Budownictwo i Inżynieria Środowiska”, 2015, nr 6.
13. Dramiński M.*, Algorytm indukcji reguł decyzyjnych w problemach klasyfikacji i wyboru cech w zadaniach wysokowymiarowych,* Rozprawa doktorska, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa 2007.

14. Duraj A., *Wykrywanie wyjątków przy użyciu wektorów nośnych,* Zeszyty Naukowe WSInf, 2017, vol 16, nr 1.
15. Durand D., *Risk Elements in Consumer Instatement Financing,* National Bureau of Economic Research, New York, 1941.
16. Freund Y., Schapire R.E, *A decision-theoretic generalization of on-line learning ana an application to boosting,* "Journal of Computer and System Sciences", 1996, nr 55.
17. Finney D.J., *Probit Analysis. 3rd Edition,* Cambridge University Press, Cambridge 1971.
18. Fisher R.A., *The use of multiple measurements in taxonomic problems,* "Annual Eugenics", 1936, nr 7.
19. Gąska D., *Zastosowanie metody SVM do oceny ryzyka bankructwa i prognozowania upadłości przedsiębiorstw, „*Śląski Przegląd Statystyczny", 2013, vol. 11, nr 17.
20. Gałda K., *Zastosowanie algorytmów genetycznych do optymalizacji modelu SVM procesu stalowniczego* , Praca magisterska, Politechnika Śląska, Katowice 2009.
21. Geisser S., *The Predictive Sample Reuse Method with Applications*,"Journal of the American Statistical Associtation", 1975, vol. 70.
22. Geron A., *Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow,* Wydawnictwo Helion, Gliwice 2018.
23. Gestel T., Baesens B., *Credit Risk Management. Basic Concepts: financial risk components, rating analysis, models, economic and regulatory capital,* Oxford University Press, New York 2009.
24. Giemza J., Zwierzchowska K., *Wprowadzenie do modelu regresji logistycznej wraz z przykładem zastosowania w pakiecie statystycznym R do danych o pacjentach po przeszczepie nerk*i, praca licencjacka, Uniwersytet Warszawski Wydział Matematyki, Informatyki i Mechaniki, 2011.
25. Goszczyński J., *Klasyfikacja tektur za pomocą SVM – maszyny wektorów wspierających,* „Inżynieria Rolnicza", 2006, nr 13.
26. Górska R., Staszkiewicz P.*, Zastosowanie algorytmu lasów losowych do prognozowania modyfikacji opinii biegłego rewidenta,* „Zarządzanie i Finanse - Journal of Management and Finance", 2017, vol. 15, nr 3.
27. Grochowiecki T., *Zastosowanie regresji logistycznej do identyfikacji czynników ryzyka wystąpienia powikłań pooperacyjnych po jednoczesnej transplantacji trzustki i nerki,* StatSoft Polska, 2011.

28. Hand D.J., Henley W.E, *Statistical Classification Methods in Consumer Credit Scoring: a Review*, "Journal of the Royal Statistical Society", 1997, Part 3.
29. Hołda A., *Wykorzystywanie drzew decyzyjnych w prognozowaniu upadłości przedsiębiorstw w branży budowlanej*, Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie, 2009, nr 796.
30. Ignaciuk G., *Zastosowanie metod scoringowych w działalności bankowej,* Statsoft, 2010.
31. Jackowska B., *Efekty interakcji między zmiennymi objaśniającymi w modelu logitowym w analizie zróżnicowania ryzyka* zgonu, „Przegląd Statystyczny", 2011, R. LVII, Zeszyt 1-2.
32. Jakubczyk-Gałczyńska A., Kristowski A., Jankowski R., *Idea zastosowania sztucznej inteligencji w prognozowaniu wpływu drgań komunikacyjnych na odpowiedź dynamiczną budynków mieszkalnych*, XI Konferencja Nowe Kierunki Rozwoju Mechaniki, Sarbinowo Morskie 2015.
33. Janc A., Kraska M., *Credit-scoring. Nowoczesna metoda oceny zdolności kredytowej*, Bibloteka Menadżera i Bankowca, Warszawa 2001.
34. Jankowski S., *Statystyczne systemy uczące się – modelowanie i klasyfikacja*, Materiały do wykładu i projektu Sieci neuronowe i neurokomputery, 2012.
35. Jurkiewicz H., *Nieeuklidesowe sieci neuronowe*, Praca magisterska, Uniwersytet Mikołaja Kopernika Wydział Fizyki, Astronomii i Informatyki Stosowanej Katedra Informatyki Stosowanej, Toruń 2009.
36. Koronacki J., Ćwik J., *Statystyczne systemy uczące się,* Akademicka Oficyna Wydawnicza EXIT, Warszawa 2015.
37. Kmieć D., *Zastosowanie modelu logitowego do analizy czynników wpływających na bezrobocie wśród ludności wiejskiej*, Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego Ekonomika i Organizacja Gospodarki Żywnościowej, 2015, nr 110.
38. Kozak J., Juszczyk P., *Algorytmy do konstruowania drzew decyzyjnych w przewidywaniu skuteczności kampanii telemarketingowej banku,* „Studia Informatica Pomerania", 2016, nr 1.
39. Kuchciński A., *Ryzyko kredytowe w działalności* banku, Kwartalnik Naukowy Uczelni Vistula, 2016, nr 2(48).
40. Łapczyński M., *Drzewa klasyfikacyjne w badaniach satysfakcji i lojalności klientów*, Statsoft Polska, 2003.

41. Marcinkowska J., *Metody statystyczne i eksploracji danych (data mining) w ocenie występowania omdleń w grupie częstoskurczu z wąskim zespołem QRS (AVNRT i AVRT*, Rozprawa doktorska, Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu, 2015.
42. Marcinkowska-Ochtyra A., *Ocena przydatności obrazów hiperspektralnych APEX oraz maszyn wektorów nośnych (SVM) do klasyfikacji roślinności subalpejskiej i alpejskiej Karkonoszy,* Rozprawa doktorska, Uniwersytet Warszawski, 2016.
43. Matuszczyk A., *Credit scoring,* Wydawnictwo CeDeWu, Warszawa 2018.
44. Matuszczyk A.*, Dotychczasowe oraz nowe trendy w metodzie "credit scoring",* Zeszyty Ekonomiczne Uniwersytetu Szczecińskiego, 2009, nr 548.
45. Matuszczyk A., *Ryzyko kredytowe*, w: „Zarzadzanie ryzykiem w banku komercyjnym", pod red. M. Iwanicz-Drozdowksa, Wydawnictwo Poltext, Warszawa 2017.
46. Matczak E., Kozłowski W., *Zastosowanie metody drzew klasyfikacyjnych w analizie aspiracji edukacyjnych rodziców*, Instytut Badań Edukacyjnych, Warszawa 2001.
47. McCulloch W. S., Pitts W., *A Logical Calculus of the Ideas Immanent in Nervous Activity*, "The Bulletin of Mathemical Biophysics", 1943, vol. 5, no. 4.
48. Migut G., Wątroba J., *Skoring kredytowy a modele data mining,* „Rynek terminowy", 2015, nr 1.
49. Misztal M., *Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 2014, nr 328.
50. Odrzywołek K., *Wykorzystanie głębokich sieci neuronowych w weryfikacji mówcy (Deep neural networks in speaker recognition)*, Praca magisterska, Akdemia Górniczo-hutnicza im. Stanisława Staszica w Krakowie, 2016.
51. Pisula T., *Metodyczne aspekty zastosowania modeli skoringowych do oceny zdolności kredytowej z wykorzystaniem metod ilościowych,* Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 2013, nr 33.
52. Płoński P., *Zastosowanie wybranych metod przekształcania i selekcji danych oraz konstrukcji cech w zadaniach klasyfikacji i klasteryzacji,* Rozprawa doktorska, Politechnika Warszawska, Warszawa 2016.
53. Polak A. J., Mroczka J., *Regularyzacja identyfikacji obiektów złożonych opisanych modelami nieliniowymi*, „Pomiary Automatyka Kontrola", 2007, vol. 53, nr 9.

54. Prokopowicz D., *Główne determinanty zastosowania metody credit scoring w zarządzaniu ryzykiem kredytowym*, Kwartalnik Naukowy Uczelni Vistula, 2015, nr 1(43).
55. Protasiewicz J., *Zastosowanie sieci neuronowych do analizy rynku energii elektrycznej w Polsce*, Rozprawa doktorska, Instytut Badań Systemowych Polskiej Akademii Nauk, Warszawa 2008.
56. Provost F., Fewcett T., *Analiza danych w biznesie. Sztuka podejmowania skutecznych decyzji*, Wydawnictwo Helion, Gliwice 2015.
57. Przanowski K., *Credit scoring. Studia przypadków procesów* biznesowych, Oficyna Wydawnicza Szkoła Główna Handlowa w Warszawie, Warszawa 2015.
58. Przywara D., *Drzewa decyzyjne, metody budowania, zastosowania*, Praca zaliczeniowa do kursu „Informatyka systemów autonomicznych", Wydział Elektroniki Politechniki Wrocławskiej, Wrocław 2007.
59. Quinlan J. R., *Induction of decision trees*, „Machine Learning Journal", 1986, vol. 1,.
60. Raczkiewicz D., *Zastosowanie analizy regresji w reprezentacyjnych badaniach społeczno-gospodarczych*, „Econometrics", 2016, nr 1(51).
61. Rosenblatt F., *The perceptron: a probabilistic model for information*, "Psychological Review", 1958.
62. Rusak O., *Ryzyko kredytowe jako jedno z ryzyk w działalności banku*, Zeszyty Naukowe Uniwersytetu Przyrodniczo-humanistycznego w Siedlcach, 2015, nr 104.
63. Sączewska-Piotrowska A., *Zastosowanie krzywych ROC w analizie ubóstwa miejskich i wiejskich gospodarstw domowych*, „Przegląd Statystyczny, 2016, nr 2.
64. Si S., Zhang H., Keerthi S. S., Mahajan D., Dhillon I. S., Hsieh C. J., *Gradient Boosted Decision Trees for High Dimensional Sparse Outpu*t. In ICML, 2017.
65. Siderska J., *Analiza możliwości zastosowania sieci neuronowych do modelowania wartości kapitału społecznego w firmach IT*, „Economics and Management", 2013, nr 1.
66. Krawiec K., Stefanowski J., *Uczenie maszynowe i sieci neuronowe*, Wydanie II, Wydawnictwo Politechniki Poznańskiej, Poznań 2004.
67. Skrobała A., *Zastosowanie sztucznych sieci neuronowych do optymalizacji rozkładów dawek w radioterapii stereotaktycznej obszarów wewnątrzczaszkowych*, Rozprawa doktorska, Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu, 2015.
68. Staniszewska A., *Zarządzanie portfelem kredytowym banku*, Oficyna Wydawnicza Szkoła Główna Handlowa w Warszawie, Warszawa 2012.

69. Stanton T.H., „Credit Scoring and Loan Scoring: Tools for Improved Management of Federal Credit Programs", Grant Report, July 1999.
70. Świątczak E., *Badanie zdolności kredytowej przedsiębiorstwa jako sposób na ograniczenie ryzyka kredytowego banku w procesie kredytowania przedsiębiorstw*, Zeszyty Studenckie Wydziału Ekonomicznego Uniwersytetu Gdańskiego „Nasze Studia", 2009, nr 4.
71. Szeliga M., *Data science i uczenie* maszynowe, Wydawnictwo Naukowe PWN, Warszawa 2017.
72. Thomas L.C., „A Survey of Credit and Behavioural Scoring: Forecasting financial risk of lending to consumers", University of Edinburgh, UK, Credit Research Centre Working Papers No. 99/2.
73. Tłuczak A., *Zastosowanie dyskryminacyjnych modeli przewidywania bankructwa do oceny ryzyka upadłości* przedsiębiorstw, Zeszyty Naukowe Wyższej Szkoły Bankowej we Wrocławiu, 2013, nr 2(34).
74. Tyce M., *Drzewa decyzyjne z użyciem pakietu R. Zastosowanie w badaniach występowania nawrotu choroby u pacjentek z nowotworem piersi,* praca licencjacka, Uniwersytet Warszawski Wydział Matematyki, Informatyki i Mechaniki, 2011.
75. Vapnik V.V., Chervonenkis A.Y., Theory of pattern recognition, "Nauka", 1974, nr 107.
76. Walesiak M., Gatnar E., *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2012.
77. Wątroba J., *Skoring kredytowy a modele data mining*, StatSoft, 2004.
78. Wiatr M. S., Jagiełło R., *Ryzyko kredytowe*, [w:] *Współczesna bankowość*, red. M. Zaleska, Difin, Warszawa 2007.
79. Wilczewski G., *InTrees: Modularne podejście do Drzew Decyzyjnych*, praca magisterska, Uniwersytet Mikołaja Kopernika Wydział Matematyki i Informatyki, Toruń 2008.
80. Woldańska M., *Zastosowanie drzew klasyfikacyjnych w badaniu zjawiska migracji klienta sieci telefonii komórkowej,* praca dyplomowa inżynierska, Politechnika Warszawska Wydział Elektroniki i Technik Informacyjnych Instytut Informatyki, 2013.
81. Wójciak M., *Metody oceny ryzyka kredytowego,* Polskie Wydawnictwo Ekonomiczne, Warszawa 2007.

82. Wójcik F., *Prognozowanie dziennych obrotów przedsiębiorstwa za pomocą algorytmu XGBoost – studium przypadków,* Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, 2018, nr 375.

Źródła internetowe

1. Blog matematyczny, https://mathspace.pl/matematyka/receiver-operating-characteristic-krzywa-roc-czyli-ocena-jakosci-klasyfikacji-czesc-7/.
2. Dokumentacja algorytmów uczenia maszynowego, https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html.
3. Encyklopedia internetowa, http://encyklopedia.naukowy.pl/Walidacja_krzy%C5%BCowa.
4. Kurs modelowania ryzyka kredytowego w języku Python, https://365datascience.com/courses/credit-risk-modeling-in-python/.
5. Legal Information Institute, https://www.law.cornell.edu/cfr/text/12/202.2.
6. Materiały edukacyjne AGH http://galaxy.uci.agh.edu.pl/~vlsi/AI/wstep/.
7. Materiały edukacyjne Katedry Cybernetyki i Robotyki Politechniki Wrocławskiej, https://kcir.pwr.edu.pl/~witold/ai/ml_nndeeps.pdf.
8. Materiały edukacyjne Katedry Inżynierii Komputerowej Uniwersytetu Rzeszowskiego, http://www.neurosoft.edu.pl/media/pdf/jbartman/sztuczna_inteligencja/NTI%20cwiczenie_5.pdf.
9. Materiały edukacyjne Katedry Inżynierii Komputerowej Uniwersytetu Rzeszowskiego, *http://www.neurosoft.edu.pl/media/pdf/tkwater/sztuczna_inteligencja/2_alg_ucz_ssn.pdf.*
10. Miesięcznik dla specjalistów z branży IT, http://www.it-professional.pl/stacje-robocze/artykul,8536,automatyczne-uczenie-maszynowe-z-usluga-azure-ml.html.
11. „RHOL Credit Scoring" http://www.rental-housing.com/fcra/PDF/scoring.pdf.
12. Strona internetowa Machine Learning i Data Science, https://mlinpl.pl/walidacja-krzyzowa-kroswalidacja/.

Spis tabel

Spis rysunków

Spis wykresów