Studium Magisterskie

Kierunek Metody Ilościowe w Ekonomii i Systemy Informacyjne

Dzmitry Fiodarau

Nr albumu:df81755

Ocena skuteczności klasycznych metod statystycznych i technik machine learning w predykcji LGD

Praca magisterska
pod kierunkiem naukowym

dr. Zuzanna Wośko

Instytut Ekonometrii

Spis treści

Wstęp

Od momentu powstania pierwszych instytucji finansowych zagrożenie niewywiązania się dłużników ze swoich zobowiązań w ustalonym terminie stanowiło główny element działalności kredytowej. Przez stulecia wiele starożytnych banków opierało swój model biznesowy na innych źródłach dochodu, a decyzje kredytowe podejmowano w dużej mierze na podstawie przesłanek subiektywnych. Dopiero w zeszłym wieku, rozwój zaawansowanych metod analitycznych pozwolił wielu bankom, skutecznie minimalizować straty wynikające z niespłaconych zobowiązań, a także optymalizować politykę kredytową, maksymalizując zyski.

Współczesna bankowość opiera się na zaawansowanych modelach matematycznych i statystycznych, które umożliwiają instytucjom finansowym na zarządzanie ryzykiem niewypłacalności klientów. Przemiany te wynikają z jednej strony z rosnącej różnorodności instrumentów będących przedmiotem obrotu na rynkach finansowych, a z drugiej strony z opracowania przez naukowców nowych, doskonalszych metod służących do opisu ryzyka oraz zarządzania nim.[1]

Upadek niemieckiego Herstatt Banku w 1974 roku[2], spowodowany w dużej mierze stratami na rynku walutowym, spowodował powstanie Bazylejskiego Komitetu Nadzoru Bankowego, celem którego było zapewnienie stabilności systemu finansowego. Bazylea II została przyjęta w 2004 roku, główną ideą której było rozszerzenie wcześniejszych regulacji. Istotny aspekt nowej regulacji, polegał na wdrożeniu banków do stosowania metod ratingowych do wyznaczania wymogów kapitałowych.

Globalny kryzys na rynku hipotek w USA w 2008 roku w dużej mierze był przyczyną przyjęcia Bazylei III, głównym zadaniem której było zwiększenie odporności banków poprzez nałożenie surowych wymogów płynnościowych, dodatkowych restrykcji kapitałowych oraz wprowadzenie nowego mechanizmu bufora kapitałowego.

Obecnie każdy bank w Unii Europejskiej zobowiązany jest do funkcjonowania zgodnie z regulacjami Bazylejskimi odpowiednio dostosowując swoją politykę zarządzania ryzykiem finansowym. Choć regulacje Bazylejskie mają charakter zaleceń, w wielu krajach, zostały wprowadzone jako obowiązkowe. Ponadto, krajowe organy regulacyjne są uprawnieni do monitorowania działalności banków, takie jak Komisja Nadzoru Finansowego (KNF) w Polsce, które dbają o zgodność polityki banków z aktualnymi zasadami.

Wgląd w historyczne aspekty regulacji wskazuję, jak ważną rolę one odgrywają w procesie zarządzania ryzykiem, co prowadzi do wniosku o trudności wdrażania nowych metod. Obecnie banki oraz inne instytucje finansowe wykorzystują zarówno klasyczne metody statystyczne, jak i algorytmy uczenia maszynowego do oceny ekspozycji portfeli kredytowych. Natomiast metody deep learning nie są stosowane, ponieważ dotychczasowe przepisy nie pozwalają na podejmowanie decyzji kredytowych w oparciu o modeli, których sposób uczenia nie jest interpretowalny i logiczny.

Główną hipotezą badawczą niniejszej pracy to założenie, że modele typu „black box" takie jak XGBoost oraz sieci neuronowe osiągają wyższą trafność predykcyjną w zakresie modelowania LGD i jego składowych niż modele regresji logistycznej.

---

[1] Wojciech Kuryłek, Modelowanie ryzyka portfela kredytowego. Część I, str. 1

[2] https://www.knf.gov.pl/dla_rynku/pakiet_crd4/historia_zalozenia#:~:text=%C5%B9r%C3%B3d%C5%82o%20regulacji%20bazylejskich%20si%C4%99ga%201974,marki%20niemieckiej%20i%20dolar%C3%B3w%20ameryka%C5%84skich (dostęp 01.04.2025)

Rozdział I. Modelowanie przewidywanej straty w kontekście ryzyka kredytowego

1.1 Definicja i komponenty ryzyka kredytowego

Ponieważ istnieje oczywista asymetria informacyjna między bankami a kredytobiorcami, bank udzielając kredytu ponosi ryzyko, gdyż nie ma pewności w momencie podpisania umowy, że należność zostanie w całości i terminowo spłacona. Ten rodzaj ryzyka nazywamy ryzykiem kredytowym. Jest on ściśle związany z jakością kredytobiorców korzystających z usług banku.[3]

Choć istnieje wiele różnych definicji ryzyka kredytowego, w niniejszej pracy przyjęto definicję Komisji Nadzoru Finansowego (KNF), będącą organem nadzoru nad rynkiem finansowym w Polsce. Według KNF, ryzyko kredytowe, to ryzyko nieoczekiwanego niewykonania zobowiązania lub pogorszenia się zdolności kredytowej zagrażającej wykonaniu zobowiązania.[4] Dotyczy to zarówno osób fizycznych, jak i podmiotów gospodarczych. Warto rozróżniać pojęcia ryzyka kredytowego i zdolności kredytowej Ryzyko kredytowe obciąża udzielającego kredytu i istnieję w każdej transakcji kredytowej. Z kolei zdolność kredytowa stanowi cechę dłużnika, określającą jego wypłacalność .

Zgodnie z podejściem IRB (*ang. Internal Ratings-Based Approach*) , który został określony w ramach regulacji Bazylea II oraz Bazylea III, ważnym wskaźnikiem jest oczekiwana strata (*ang. expected loss, EL*). Wkaźnik przedstawia średnią wartość strat, jakie instytucja finansowa może ponieść w związku z ekspozycją kredytową w danym okresie.[5,6]

Wzór ten przyjmuje postać:

$$EL = PD * EAD * LGD \tag{1}$$

gdzie:

PD (*Probability of Default*) – prawdopodobieństwo niewypłacalności zobowiązania

EAD (*Exposure at Default*) – ekspozycja kredytowa w momencie niewypłacalności

LGD (*Loss Given Default*) – procentowy udział straty poniesionej przez instytucję w przypadku niewypłacalności strony zobowiązanej.

Zaawansowane podeście zakłada modelowanie tych trzech parametrów i nie jest powszechnie stosowane. Wiele mniejszych instytucji finansowych posługują się podejściem standardowym (*ang. Standardized Approach*). Podejście to polega przyjęciu ustalonych wartości parametrów, które są dostarczane przez KNF.

Niezależnie od podejścia, średni poziom strat, jakiego instytucja finansowa może się spodziewać przy udzielaniu kredytów, jest standardowym rodzajem strat poniesionych przez instytucję finansową. Zgodnie z poleceniami Bazylei II i III, wartość oczekiwanej straty musi być pokrywana przez bank wydzielonym odpowiednio kapitałem. Kadra zarządzająca organizacją podejmuję decyzji oraz określa cele strategiczne w oparciu o wielkość oczekiwanej straty.

---

[3] Wojciech Kuryłek, Modelowanie ryzyka portfela kredytowego. Część I, str. 6

[4] Komisja Nadzoru Finansowego, *BION w bankach – mapa klas ryzyka i ich definicje*, KNF. str. 1

[5] Basel Committee on Banking Supervision, 2005, s. 8

[6] Basel Committee on Banking Supervision. (2017). *Basel III: Finalising post-crisis reforms*. Bank for International Settlements, s. 56.

Natomiast dla straty nieoczekiwanej nie istnieje powszechnie przyjętego wzoru. W podejściu zaawansowanym stosowany jest stosowany wzór polegający na oszacowaniu wagi ryzyka za pomocą funkcji odwrotnej rozkładu normalnego, wykorzystywana do modelowania strat przy poziomie ufności 0.999.

Wzór wygląda następująco:

$$\mathrm{UL} = \mathrm{RW} * \mathrm{EAD} = \left(\mathrm{LGD} \cdot \mathrm{N}\left(\frac{1}{\sqrt{1-\mathrm{R}}} \cdot \mathrm{G}(\mathrm{PD}) + \sqrt{\frac{\mathrm{R}}{1-\mathrm{R}}} \cdot \mathrm{G}(0.999) \right) - \mathrm{LGD} \cdot \mathrm{PD} \right) * \mathrm{A}' * 12.5 * 1.06^{7} * \mathrm{EAD} \quad (2)$$

gdzie:

RW- waga ryzyka

G-Odwrotność dystrybuanty rozkładu normalnego

R-Współczynnik korelacji aktywów

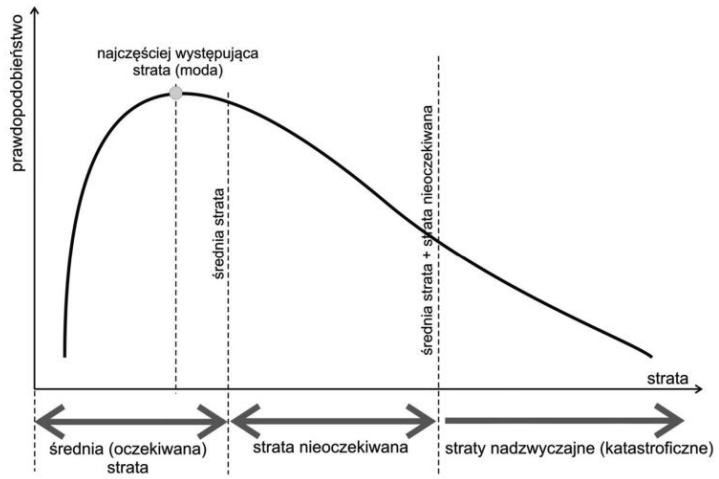
A-Współczynnik zapadalności

12.5-aktywa ważone ryzykiem

1.06-współczynnik korekcyjny wprowadzony przez Bazyleę II

Rysunek 1 przedstawia rozkład prawdopodobieństwa strat finansowych. Straty dzielą się na trzy grupy: oczekiwane, nieoczekiwane oraz nadzwyczajne. Banki są zobowiązane utrzymywać rezerwy na pokrycie strat oczekiwanych oraz nieoczekiwanych. Straty nadzwyczajne są rzadkie, w związku z tym nie ma bezwzględnego obowiązku pokrycia ich kapitałem. Z kolei Bazylea III wprowadziła mechanizmy buforów kapitałowych, które mają na celu zwiększać odporność banku w odniesieniu strat nadzwyczajnych.

Rysunek 1.1 Rozkład strat w zależności od prawdopodobieństwa ich wystąpienia[8]



Źródło: Wiszniowski E., Model szacowania utraty wartości instrumentów finansowych w założeniach MSSF 9 – rachunkowość czy inżynieria finansowa?, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 313, 2013, s. 170–188.

[7] Komisja Nadzoru Finansowego, *Zarządzanie ryzykiem modeli w zakresie działalności podmiotów sektora bankowego ze szczególnym uwzględnieniem modeli ryzyka kredytowego – podstawowe zagadnienia*, Paweł Grodź, Bartosz Lewandowski, Kamil Simka, Maja Tuszyńska, Warszawa 2024. s. 53

[8] Wiszniowski E., *Model szacowania utraty wartości instrumentów finansowych w założeniach MSSF 9 – rachunkowość czy inżynieria finansowa?*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 313, 2013, s. 170–188.

Warto zwrócić uwagę na definicję niewypłacalności(*ang. Definiton of Default)*. Historycznie Definicja niewypłacalności klienta różniła się w zależności od banku . Jednakże Europejski Urząd Nadzoru Bankowego (*EBA*) wprowadził jednolite wymogi, zgodnie z którymi klient uznawany jest za niewypłacalnego, jeżeli opóźnienie w spłacie zobowiązania przekracza 90 dni i spełniają się dwa warunki jednocześnie: zaległość musi przekraczać 100 euro dla klienta detalicznego lub 500 euro dla klienta korporacyjnego oraz zaległa kwota musi stanowić co najmniej 1% całkowitej ekspozycji klienta.[9] Zaczynając od tego momentu, klient może zostać uznany za niewypłacalnego. Poprawki mieli na celu ustandaryzowanie podejścia definicji niewypłacalności we wszystkich instytucjach finansowych w UE, niemniej jednak w praktyce istnieje istotne rozbieżności w jej zastosowaniu.

1.2 Modelowanie parametru LGD

Parametr LGD (*ang*. Loss Given Default) albo strata z tytułu niewykonania zobowiązania, jest jednym z centralnych parametrów podczas oceny ryzyka kredytowego. Zgodnie ze wzorem 1.1, parametr jest jednym z trzech elementów wchodzących w skład wzoru na straty oczekiwanej (*ang. Expected loss*). LGD przedstawia stosunek straty na ekspozycji, która nastąpiła w rezultacie niewykonania zobowiązania do kwoty należności i mierzy stratę w sensie ekonomicznym, nie księgowym. Istnieje dwa rodzaje modeli LGD:

- zasadniczym celem modeli non-default dla ekspozycji pracujących (*ang. performing exposures*) jest przewidzenie stosunku straty ekonomicznej do kwoty, jaka pozostała do spłaty zobowiązania kredytowego w hipotetycznym momencie niewykonania zobowiązania. Zatem przyjmuję się, że zobowiązanie kredytowe nie zostanie zrealizowane.
- celem modeli in-default dla portfela niewykonanych zobowiązań (*ang. defaulted exposures*) jest określenie różnicy pomiędzy stratą ekonomiczną, a bieżącą kwotą pozostającej do spłaty każdego niewykonanego zobowiązania kredytowego.

Warto zaznaczyć, że w zależności od rodzaju modeli LGD , różnią się koncepcji i wzory matematyczne.

W sytuacji gdy klient w niniejszym momencie znajduję się w sytuacji niewykonania zobowiązania(*ang. in-default*), parametr LGD stanowi odwrotność stopy odzysku z ekspozycji(*ang*. *Recovery Rate*) i wygląda następująco:

$$LGD = 1 - RR = 1 - \frac{\sum_{t=1}^{T} \frac{CF_t}{(1+r)^t}}{EAD} \tag{3}$$

gdzie:

EAD – wartość ekspozycji na datę niewykonania zobowiązania

CF – przepływy pieniężne

r – stopa dyskontująca

[9] European Banking Authority (EBA), *Guidelines on the application of the definition of default under Article 178 of Regulation (EU) No 575/2013*, Final Report, EBA/GL/2016/07.

Natomiast dla warunku, gdy klient jeszcze nie w jest momencie niewykonania zobowiązania i rozważany hipotetyczny moment jego niewykonania, parametr LGD można przedstawić w dwuczynnikowej postaci:

$$LGD = (1 - \mathrm{PC}) * (1 - \mathrm{RR}) \quad (4)$$

gdzie:

PC - prawdopodobieństwo uzdrowienia, tj. wykonanie wszystkich powstałych dotychczas zobowiązań,

RR - stopa odzysku niewykonanych zobowiązań kalkulowana od wartości ekspozycji z ostatniego momentu wejścia do stanu niewykonania zobowiązania.

Dla Wzoru 3 możemy rozważyć modelowanie oddzielnie każdego czynnika.

W praktyce bankowej przyjmuje się również rozróżnienie na dwa typy LGD: zrealizowaną (*ang. realized LGD*) oraz obserwowaną (*ang. observed LGD*).

Obserwowana wartość LGD to średnia ważona zrealizowanych wartości LGD dla ekspozycji, które w tym momencie zostały rozwiązane lub pozostają nierozwiązane, ale dobiegł końca maksymalny okres odpracowania. Funkcją obserwowanej wartość LGD jest pokazanie rzeczywistego poziomu strat historycznych ponoszonych przez bank w przypadku niewywiązania się klientów z zobowiązań.

LGD zrealizowana to realna strata, jaką ponosi bank w momencie zakończenia spłacania zobowiązania, a proces odzyskiwania należności został zakończony. W większości przypadków w bankach ustalany okres workout, rozumiany jako czas, jaki upłynął od daty niespłacenia do ukończenia procesu odzysku należności. Okres ten nie ma regulacyjnie ustalonych kryteriów, z tego powodu może różnić w zależności od instytucji finansowej oraz konkretnej sytuacji. Zrealizowaną LGD obliczana jako stosunek rzeczywistej straty do wartości w momencie niewykonania zobowiązania[10]:

$$LGD_{Zrealizowane} = \frac{Zrealizowana\ strata}{Zrealizowane\ EAD} \quad (5)$$

W ujęciu praktycznym, dla obliczenia zrealizowanej straty, rozważane są: niespłacone zobowiązania, odzyskane środki, koszty windykacji oraz brana pod uwagę wartość pieniądza

w czasie. Wzor wygląda następująco: w czasie. Wzor wygląda następująco:

$$LGD_{Zrealizowane} = \frac{EAD - (\sum_{j=1}^{n} NPV(odzyski) - \sum_{j=1}^{n} NPV(koszty\ bezpośrednie) - \sum_{j=1}^{n} NPV(koszty\ pośrednie)}{EAD} \quad (6)$$

gdzie:

EAD- wartość ekspozycji na datę niewykonania zobowiązania

NPV(odzyski) – wartość bieżąca odzyskanych środków

NPV(koszty bezpośrednie) – koszty poniesione bezpośrednio w związku z działaniami egzekucyjnymi(komornik)

NPV(koszty pośrednie) –nakłady operacyjne banku związane z procesem odzysku należności(postępowania administracyjne)

[10] https://www.openriskmanual.org/wiki/Realised_LGD (dostęp 06.04.2025)

n – liczba przepływów

Wszystkie odzyskane środki, zgodnie z wytycznymi Europejskiego Urzędu Nadzoru Bankowego[11] i rekomendacją KNF, muszą być obliczone jako wartość bieżącą netto (NPV). Głównym powodem jest zmienność wartości pieniądza w czasie. Wzór przyjmuje postać:

$$Roczna\ stopa\ dyskontowa = Stopa\ referencyjna\ KNF + 5\% \quad (7)$$

W ramach metody IRB, instytucje finansowe są zobowiązani estymować średnie długoterminowe LGD (*ang. Long-Run Average LGD*), będące elementem kalibracji modeli. Celem tej metryki jest ilustracja przeciętnej wartości LGD, która występuje w długim okresie oraz pokazuję bezwzględną wartość, niezależną warunków gospodarczych.

W celu wyznaczenia średniej długoterminowej LGD, analizowany zbiór danych powinien obejmować okres pięcioletni, natomiast dla wszystkich rodzajów przedsiębiorstw-siedmioletni. Podstawą danych są: zrealizowane wartości LGD dla rozwiązanych przypadków niewykonania zobowiązania, zrealizowane wartości LGD dla nierozwiązanych przypadków niewykonania zobowiązania oraz zrealizowana wartość LGD dla niewykonanych zobowiązań.

Modele powinny zostać dopasowane w taki sposób, żeby przewidywane wartości LGD nie były średnio niższe niż wartości historyczne średniej długoterminowej LGD. W sytuacji niespełnienia danego warunku, konieczne jest zastosowanie marginesu ostrożności MoC (*ang. Margin of Conservatism*), zabezpieczający przed niepewnością otrzymanych wyników.

Rozdział II. Metody statystyczne, predykcyjne oraz ich właściwości

2.1 Bias-Variance Trade-off

„Bias-Variance Trade-off" jest ważnym pojęciem w uczeniu maszynowym i statystyce, które pozwoli na właściwy wybór modelu do modelowania. Kompromis między obciążeniem a wariancją który polega na minimalizacja błędu predykcji całkowitego.

Rozważmy model $\hat{f}(x)$, który został wytrenowany na zbiorze treningowym Tr, oraz punkt testowy (x_0 ,y_0), pochodzący z tej samej populacji, co dane treningowe.[12]

Rzeczywistą zależność między zmiennymi można opisać równaniem:

$$Y = f(X) + \varepsilon \quad (8)$$

gdzie:

$f(X) = E[Y \mid X = x]$ to funkcja wartości oczekiwanej, pokazującą zależność zmiennej X od zmiennej Y

ε to losowy błąd

Jakość dopasowania modelu oceniono poprzez obliczenie oczekiwanego błędu kwadratowego między rzeczywistą wartością y_0 a przewidywaną wartością $\hat{f}(x_0)$:

[11] European Banking Authority, *Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures (EBA/GL/2017/16)*, 2017, s. 71–72

[12] Statistical Learning, Trevor Hastie Robert Tibshirani, Springer,2021, str 56-57

$$E\left(y_0 - \hat{f}(x_0)\right)^2 \tag{9}$$

Powyższy wzor rozłożono na składniki:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\epsilon) \tag{10}$$

gdzie:

$Var\left(\hat{f}(x_0)\right)$ - wariancja modelu, mierzy, zmienność przewidywania modelu w zależności od różnych zbiorów treningowych.

$\left[Bias\left(\hat{f}(x_0)\right)\right]^2$ – kwadrat obciążenia, mierzy, jak mocno średnia przewidywana wartość $\hat{f}(x_0)$ odbiega od prawdziwej wartości $f(x_0)$

$Var(\epsilon)$ to wariancja błędu losowego, która nie zależy od modelu.

W równaniu 10 jest pokazane, że nie można jednocześnie zminimalizować obciążenia i wariancji, ponieważ redukcja jednego ze składników prowadzi do wzrostu drugiego.

Tabela 3.1 przedstawia zależność między złożonością modelu, a jego błędem całkowitym, polegającym na kompromisie „bias-variance trade-off”:

Tabela 2.1.1 Zależność między złożonością modelu, a jego błędem całkowitym

Złożoność modelu	Obciążenie	Wariancja	Błąd całkowity
Niska	Wysokie	Niska	Wysoki
Średnia	Optymalne	Optymalna	Niski
Wysoka	Niskie	Wysokie	Wysoki

Źródło: opracowanie własne

W prostych modelach takich jak regresja liniowa lub regresja logistyczna zakładana jest liniowa zależność między zmiennymi. Dane modele charakteryzują się niską wariancją, ale wysokim obciążeniem, upraszają rzeczywistość i wychwytują złożoną strukturę danych. W konsekwencji mogą nie dopasowywać się dobrze do rzeczywistych zależności, co prowadzi do niedouczenia (*ang. underfitting*).Natomiast złożone modele, takie jak drzewa decyzyjne bez ograniczeń głębokości lub sieci neuronowe mają niskie obciążenia, ale wysoką wariancję. To oznacza, że są podatne na dopasowanie do konkretnych danych. W konsekwencji to często prowadzi do przeuczenia (*ang. overfittingu*) modelu i niskiej zdolności do przewidywania dla nowych obserwacji.

2.2 Regresja logistyczna

Regresja liniowa jak i regresja logistyczna stanowią przypadek rodzajów modeli upraszających rzeczywistość. Regresja logistyczna najczęściej stosowana w modelowaniu ryzyka kredytowego, także podczas estymacji parametru LGD.

Regresja logistyczna wchodzi do rodziny modeli dwumianowych i stanowi szczególny przypadek regresji liniowej. Analogicznie jak i modelach regresji liniowej, regresja logistyczna próbuję zbadać związek zmiennej objaśnianej za pomocą zmiennych objaśniających. Kluczowa różnica polega na tym, że zmienna zależna wykorzystuję funkcję logistyczną w celu ograniczenia jej przedziału [0,1].

Funkcja dystrybuanty rozkładu logistycznego ma postać:

$$p_i = F(x_i\beta) = \frac{e^{(x_i'\beta)}}{1 + e^{(x_i'\beta)}} = \frac{1}{1 + e^{(-x_i\beta)}} \tag{11}$$

gdzie:

p_i to prawdopodobieństwo, że zmienna zależna przyjmie wartość 1

x_i to wektor zmiennych objaśniających

β to wektor estymatorów

Funkcja jako dystrybuatna ma oczywisty kształt "krzywej typu S" albo sigmoidealny, co eliminuję problem generowania prognozowanych wartości spoza przedziału [0,1] , który jest spotykany w liniowym modelu prawdopodobieństwa. Wartości spoza przedziału [0,1] nie mają interpretacji probabilistycznej.

Następnie w celu wygodniejszego modelowania obliczono funkcję odwrotną i skorzystano ze własności logarytmu i eksponenty:

$$x_i\beta = F^{-1}(p_i) = ln(\frac{p_i}{1 - p_i}) \tag{12}$$

Bardziej efektywnym podejściem niż modelowanie p_i względem zmiennych objaśniających X, jest podejście modelować wyrażenie $ln(\frac{p_i}{1-p_i})$ jako liniową funkcję tych zmiennych. Model ten nazywa się modelem logitowym lub logistycznym. Z definicji wynika, że logit to logarytm naturalny ilorazu prawdopodobieństw przyjęcia lub nieprzyjęcia wartości 1 przez zmienną objaśnianą. Dla p_i <0.5 logit przyjmuję wartości ujemne, natomiast dla p_i>0.5 przyjmuję wartości dodatnie.[13]

W modelu logitowym estymatory β_1 β_2,…,β_k wchodzące w skład wektoru β szacuje się za pomocą metody największej wiarygodności. Celem tej metody jest znalezienie takiego wektoru β, który maksymalizuje logarytm funkcji wiarygodności, umożliwiając najlepsze dopasowanie modelu do danych.

Główną ideą budowy modelu logitowego oraz estymacji jego współczynników jest oszacowanie prawdopodobieństwa wystąpienia określonego zdarzenia. To może być prawdopodobieństwo przeżycia katastrofy Titanica lub ryzyko niewywiązania się klienta ze zobowiązań finansowych. Model logitowy znajduje szerokie zastosowanie w wielu dziedzinach, takich jak finanse, ekonomia, medycyna, analiza danych społecznych.

W kontekście modelowania prawdopodobieństwa ważnym pojęciem jest szansa. Pojęcie to wynika z funkcji odwrotnej i jest zdefiniowane jako stosunek prawdopodobieństwa wystąpienia zdarzenia do prawdopodobieństwa jego nie wystąpienia. Szansa określa, jak bardzo prawdopodobne jest wystąpienie danego zdarzenia w porównaniu do jego niewystąpienia.

Matematyczny wzór można zapisać następująco:

[13] Red. nauk.: M. Gruszczyński i in., *Mikroekonometria. Modele i metody analizy danych indywidualnych*, Warszawa: Wolters Kluwer, 2012, 1-195

$$szansa = \frac{p_i}{1 - p_i} \tag{13}$$

gdzie:

p_i to prawdopodobieństwo wystąpienia danego zdarzenia

Do porównania szans dwóch różnych kategorii w obrębie jednej zmiennej kategorycznej wykorzystuje się iloraz szans**.** Kiedy iloraz szans jest większy od 1, oznacza to, że szansa wystąpienia zdarzenia w jednej grupie jest większa niż w drugiej. Jednakże, gdy iloraz szans jest mniejszy od 1, istnieje mniejsza szansa zajścia zdarzenia w jednej grupie niż w grupie jej odniesienia.

$$Iloraz\ Szans = \frac{szansa\ zajścia\ zdarzenia\ dla\ grupy\ 1}{szansa\ zajścia\ zdarzenia\ dla\ grupy\ 2} = \frac{\frac{p_1}{1 - p_1}}{\frac{p_2}{1 - p_2}} \tag{14}$$

W kontekście regresji logistycznej, iloraz szans wyznaczany jest poprzez podniesienia estymatora β_k do potęgi liczby e (eksponenta) i w wyniku tej operacji iloraz szans ma inną interpretację w zależności od typu zmiennej objaśniającej:

- Gdy zmienna objaśniająca jest zmienną kategoryczna, iloraz szans przedstawia szansę zajścia zdarzenia dla jednej grupy w stosunku do grupy jej odniesienia.
- Jeżeli zmienna jest zmienną ciągłą, iloraz szans opisuje szansę zajścia zdarzenia przy wzroście zmiennej objaśniającej o jedną jednostkę, przy założeniu, że wszystkie inne czynniki pozostają bez zmian.

2.3 Modele uczenia maszynowego (rodzina algorytmów Gradient Boosting)

W niniejszej pracy zostaną opisane iteracyjne algorytmy gradient boosting, w szczególności zaawansowany i najbardziej skuteczny z nich algorytm XGBoost, pozwalający na wykrycie złożonych struktur pomiędzy zmiennymi.

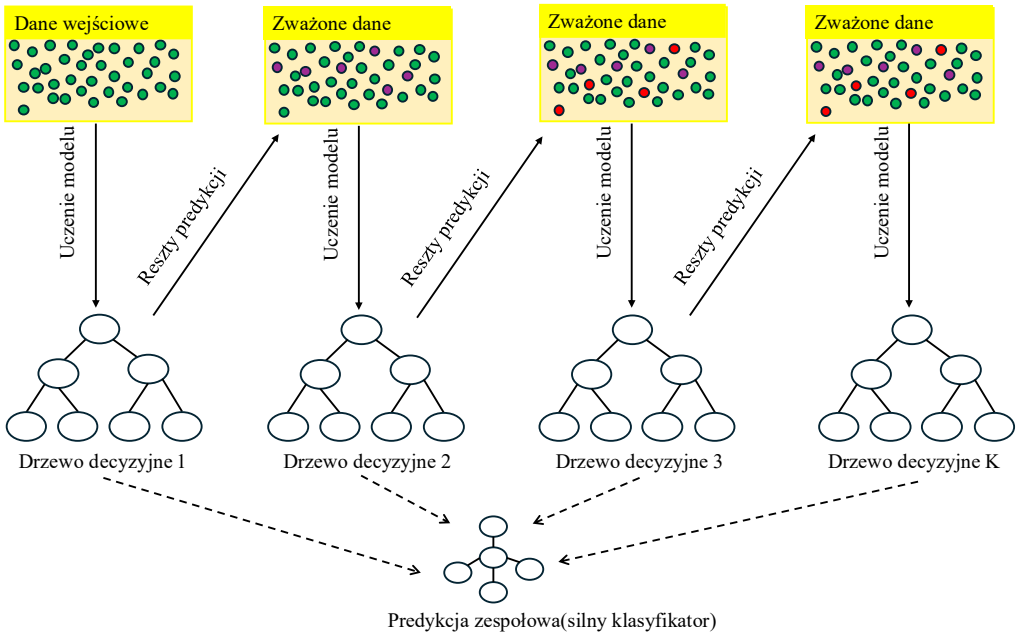
Algorytm Gradient Boosting wchodzi w skład metod uczenia zespołowego (*ang. Ensemble learning*) i został zaproponowany przez Jerome'a Friedmana w 2001 roku[14]. Celem było znalezienie algorytmu, który uczy się na podstawie wcześniejszych niepowodzeń oraz systematycznie ulepsza jakość wyników poprzez uczenie na błędach poprzednich. Natomiast algorytm XGBoost, został opracowany przez Tianqi Chena w 2016 roku[15].

Graficzny schemat działania algorytmu Gradient Boosting jest przestawiony na rysunku 2.1:

[14] Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. Annals of Statistics, 29(5), 1189–1232.

[15] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Rysunek 2.1 Schemat działania algorytmu Gradient Boosting



Źródło: opracowanie własne

W klasycznym podejściu boosting, funkcja celu jest minimalizowana w sposób iteracyjny. Funkcja celu albo funkcja kosztu zazwyczaj jest wybierana w zależności od rodzaju problemu oraz typu zmiennej objaśnianej. Algorytm opiera się na obliczaniu gradientu funkcji kosztu, z tego powodu funkcja kosztu musi być różniczkowalna względem własnej predykcji. Dla najbardziej klasycznego przypadku regresji, stosowaną funkcją kosztu jest funkcja błędu średniokwadratowego, która jest zdefiniowana jako:

$$L\big(y, F(x)\big) = \frac{1}{2}\big(y - F(x)\big)^2 \tag{15}$$

Poniższa funkcja karze duże odchylenia predykcji od rzeczywistej wartości oraz jest wypukła i różniczkowalna na całej dziedzinie.

Natomiast w procesie budowy modelu Gradient Boosting, stosuje się również inne funkcje kosztu:

- Funkcja Hubera jest skonstruowana w celu zwiększenia odporności na wartości odstające, może być stosowana dla budowy dowolnego modelu boosting, kiedy w danych jest znaczna ilość wartości odstających.
- Mając do czynienia z zmiennej zależnej zero-jedynkowej, stosuje się funkcja log-loss, która polega na estymacji funkcji logitowej.
- Do zadań klasyfikacji wieloklasowej rozszerzeniem funkcji log-loss jest entropia krzyżowa (*ang. cross-entropy*). Główną zaletą tej funkcji jest, to że mierzy różnice pomiędzy rozkładem prawdopodobieństwa rzeczywistym, a przewidywanym przez model dla różnych kategorii zmiennej zależnej.

Kolejnym krokiem jest minimalizacja wybranej funkcji kosztu. Następnie liczone jest dla każdej jednostki danych pseudo-reszty. Pseudo-reszty wyznaczają wartość, do której trzeba ulepszyć model dla uzyskania lepszej wartości predykcji w następnej iteracji. Dalej budowana jest funkcją, która szacuje wektor wartości pseudo-reszt. Ostatnim krokiem jest szukanie

optymalnej wartości współczynnika, minimalizując oszacowany wektor pseudo-reszt. Cały proces jest powtarzamy i matematycznie można zapisać następująco:

$$F_K(x) = F_0(x) + \sum_{k=1}^{K} v * \varphi_k * m_k(x) \quad (16)$$

gdzie:
$F_K(x)$ – finalny model po K iteracjach
$F_0(x)$ – model początkowy

v – współczynnik nadający wagę różnym modelom(hiperparametr)
φ_k – współczynnik nadający wagę oszacowanemu wektorowi pseudo-reszt

$m_k(x)$- oszacowany wektor pseudo-reszt

Model XGBoost przebiega w podobny sposób do tradycyjnego poniżej opisanego Gradient boosting. Procedura wygląda następująco[16]:

1. Wybór i funkcji kosztu i jej minimalizacja
2. Proces iteracyjny:

- obliczenie pseudo-reszt i Hessianu funkcji kary
- Zbudowanie drzewa decyzyjnego. Za pomocą wag i preudo-reszt tworząc optymalny podział
- Obliczenie metryki simularity score, mierząca jakość węzłów w drzewie[17]
- Ustalenie najlepszego podziału
- Stosowanie regularyzacji i hiperparametrów
- Obliczenie wyników oraz dodawanie kolejnego drzewa

3. Odświeżenie wektoru pseudo-reszt
4. Otrzymanie końcowego modelu

W Modelu XGBoost może być stosowana również regularyzacja, która pomoga regulować proces uczenia modelu.

Regularyzacja L1(*lasso*) jest techniką stosowaną w celu zmniejszenia złożoności modelu. Odbywa się to poprzez zmniejszenie wag związanych z niektórymi parametrami, co z kolei usuwa rzadkie, hałaśliwe lub nieistotne cechy z modelu. Pomaga to modelowi stać się bardziej uogólnionym i mniej podatnym na nadmierne dopasowanie.[18]

Regularyzacja L2 kara za dużą wartość podniesioną do kwadratu predykcji w liściu drzewa decyzyjnego.
Mechanizm działania jest analogiczny do Regularyzacji L1, natomiast istotną różnicą jest to, że nie może wyzerować pewne cechy z modelu. Regularyzacja L2 jest włączona do modelu XGBoost podczas liczenia wartości predykcji w konkretnym liście.

[16] https://www.geeksforgeeks.org/xgboost/ (dostęp dnia 25.04.2025)
[17] https://medium.com/@prathameshsonawane/xgboost-how-does-this-work-e1cae7c5b6cb (dostęp dnia 26.04.2025)

[18] https://fineproxy.org/pl/wiki/regularization-l1-l2/ (dostęp dnia 26.04.2025)

Podsumowując, w ramach modelowania parametru LGD, XGBoost może być skutecznym narzędziem zdolnym do modelowania niewypłacalności. XGBoost również wykrywa nieliniowe zależności w danych oraz rozpozna złożone relacje pomiędzy zmiennymi.

2.4 Modele sieci neuronowej

Subiektywne oceny oparte na doświadczeniu tradycyjnych modelach liniowych stopniowo ustępują miejsca bardziej zaawansowanym metodom oceny opartym na technikach uczenia maszynowego[19]. Modeli sieci neuronowe coraz częściej wykorzystują w praktyce bankowej.

Sztuczne sieci neuronowe są inspirowane ich biologicznymi odpowiednikami – mózgiem i układem nerwowym. Najważniejszą cechą charakterystyczną biologicznego mózgu jest jego zdolność do uczenia się i adaptacji, podczas gdy komputer nie posiada takich umiejętności. Podstawowym budulcem sieci neuronowych jest „neuron". Neuron może być postrzegany jako jednostka przetwarzania. Każdy neuron w sieci otrzymuje odpowiednio ważone informacje za pośrednictwem połączeń synaptycznych. Z neuronów, które są podłączone do niego, wytwarza wyjściową informację, przetwarzając ważoną sumę sygnałów wejściowych za pośrednictwem funkcji aktywacji.

Istnieją dwie główne kategorie architektur sieciowych rozróżniane na podstawie typu połączeń między neuronami: jednokierunkowe i rekurencyjne sieci neuronowe. Jeśli nie ma sprzężenia zwrotnego z wyjść neuronów do wejść w całej sieci, sieć neuronowa jest określana jako jednokierunkowa. W przeciwnym razie, jeśli istnieje sprzężenie zwrotne, czyli połączenie synaptyczne z wyjść do wejść (albo ich własnych wejść lub wejść innych neuronów), sieć nazywana jest rekurencyjną siecią neuronową.[20]

W niniejszej pracy, dla modelowania parametru LGD, zostanie wykorzystana architektura wielowarstwowe perceptrony (*ang. MLP*). Ten typ architektury jest jednym z najbardziej popularnych i podstawowych spośród wszystkich sieci neuronowych. Informacja w danej architekturze przepływa tylko od wejścia do wyjścia. Architektury takiego rodzaju składają się: z warstw wejściowych, warstw ukrytych oraz warstwy wyjściowej przewidującą wynik. Poniższe rysunki 2.4.1 oraz 2.4.2 przedstawiają architekturę wielowarstwowych perceptronów, złożonych z trzech segmentów dla jednej iteracji: warstwy wejściowej (*ang. Inner layer*), dwóch warstw ukrytych (*ang. Hidden layers*) oraz warstwy wyjściowej(*ang. Outer layer*). W architekturze sieciowej tego typu przepływ informacji przebiega tylko w jednym kierunku :od wejścia do wyjścia.

Na obu wykresach schematycznych, warstwy wejściowe obejmują z cztery neurony oznaczonych jako X_1, X_2, X_3 oraz X_4. Przedstawione wejściowe neurony pełnią rolę zmiennych objaśniający, odpowiednio dobranych wcześniej. Informacja z każdej zmiennej objaśniającej jest kierowana do wszystkich neuronów warstwy ukrytej za pomocą następującego wzoru:

$$z_j = \sum_{i=1}^{4} w_{ij} X_i + b_j \tag{17}$$

[19] Zhu Y., Huang Z., Zhang X., *RS-Boosting and RS-MultiBoosting: Hybrid Ensemble Machine Learning Models for SME Credit Risk Assessment in Chinese Supply Chain Finance*, Forecasting, 2022, Vol. 4, pp. 188–190.

[20] Komisja Nadzoru Finansowego, Zarządzanie ryzykiem modeli w zakresie działalności podmiotów sektora bankowego ze szczególnym uwzględnieniem modeli ryzyka kredytowego – podstawowe zagadnienia, Paweł Grodź, Bartosz Lewandowski, Kamil Simka, Maja Tuszyńska, Warszawa 2024. s. 65-66

gdzie:

z_j- wynik ważonych zmiennych objaśniających dla neuronu j

w_{ij}- wektor wag, określający wpływ X_i na neuron j

X_i- wartość zmiennej X_i

b_j- stała wartość wyrazu wolnego przypisanego neuronowi j.

Następnym krokiem, dla każdego neuronu pierwszej warstwy ukrytej, stosowana jest funkcja aktywacji ReLU (*ang. Rectified Linear Unit*):

$$a_j = ReLU(z_j) = \max(0, z_j) \tag{18}$$

Zastosowanie wskazanej funkcji pozwala wyeliminować wartości ujemne oraz zmniejsza problem zanikania gradientu.

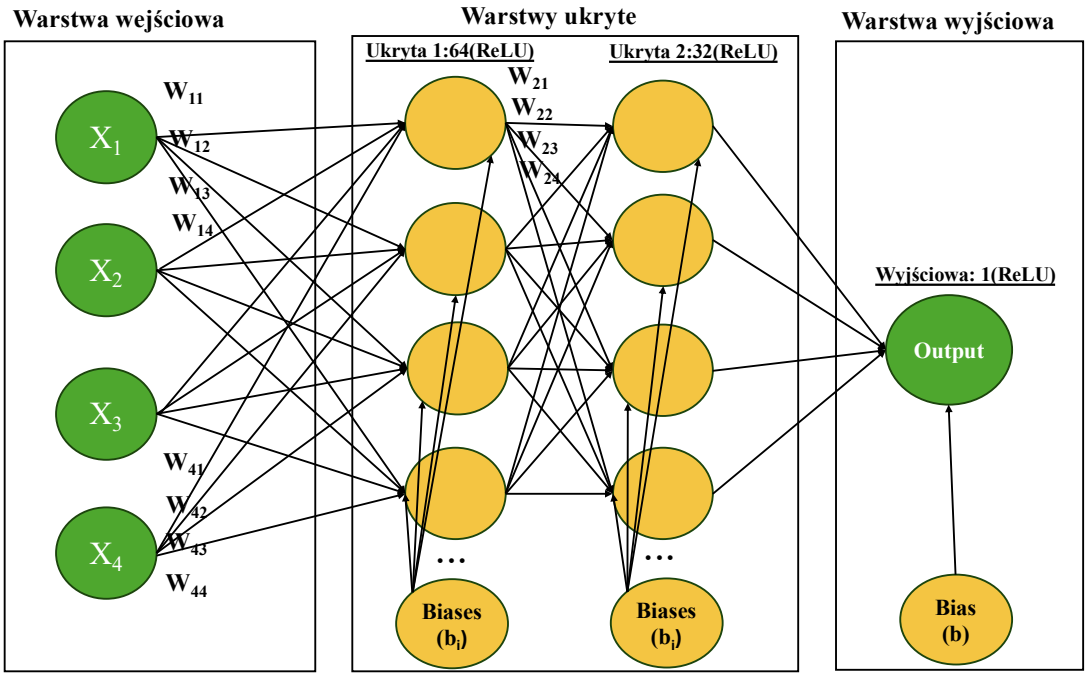
W dalszej części neurony przechodzą z pierwszej warstwy ukrytej do drugiej warstwy ukrytej. Sposób obliczenia wartości w drugiej warstwie odbywa się zgodnie z wzorem (17) i (18).

Etapem końcowym działania sieci w pierwszej iteracji jest przekazanie otrzymanych wartości z drugiej warstwy ukrytej do warstwy wyjściowej. Funkcja aktywacji w przypadku obliczania Recovery Rate określa wzór (17). Natomiast funkcją aktywacji w przypadku obliczania Probability of Cure jest sigmoid. Funkcja sigmoid jest popularną funkcją w kontekście problemów klasyfikacyjnych. Charakteryzuję się jako funkcja nieliniowa i przetwarza wartości do przedziału (0, 1). Funkcja ta ma następującą postać:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{19}$$

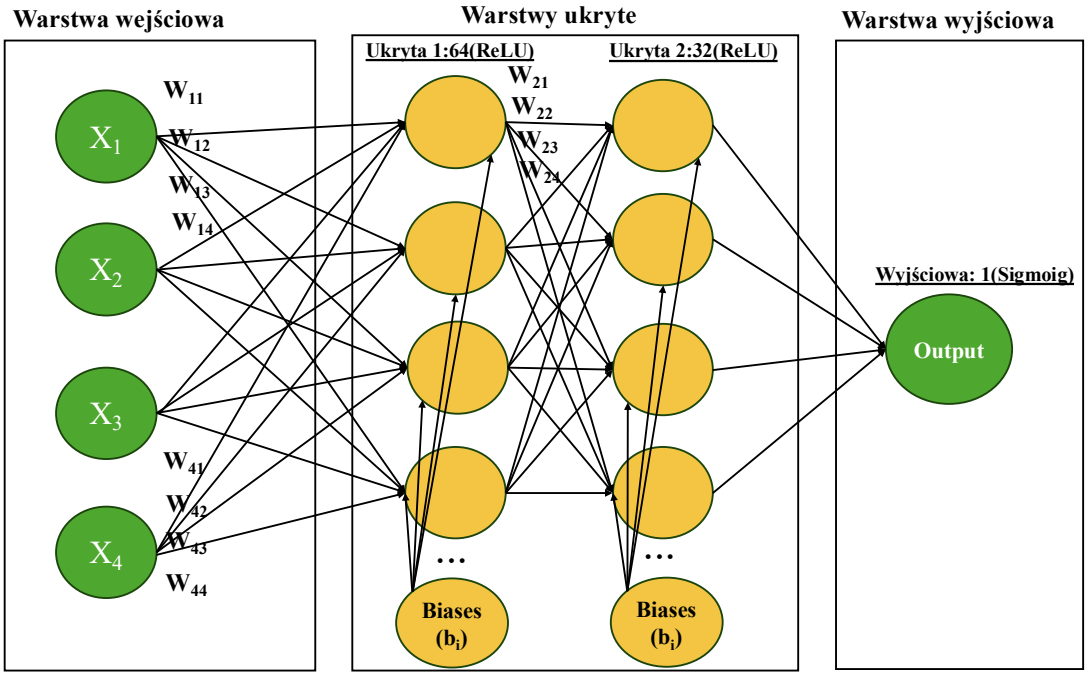
Warto podkreślić, że w przypadku zadania regresyjnego lub klasyfikacji binarnej warstwa wyjściowa zawiera tylko jeden neuron, natomiast dla klasyfikacji wieloklasowej może być ich dwa lub więcej.

Rysunek 2.4.1 Struktura wielowarstwowego perceptronu dla Recovery Rate



Źródło: opracowanie własne

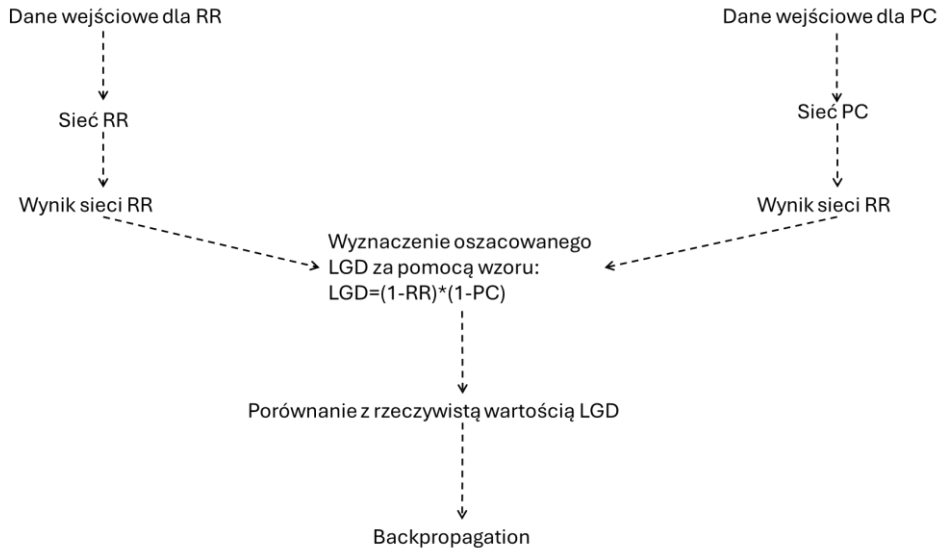
Rysunek 2.4.2 Struktura wielowarstwowego perceptronu dla Probability of Cure



Źródło: opracowanie własne

Taka architektura posiada ważną zaletę: modeluje nieliniowe zależności między zmiennymi objaśniającymi. W przeciwieństwie do metody regresji logistycznej, która podobnie zawiera funkcję sigmoid, siec neuronowa typu MLP, przechodzi przez proces iteracyjny, stopniowo ulepszając wartości wag i końcowy wynik. Jednak główną wadą takiego modelu jest problem przeuczenia. Dany problem został Szczegółowo omówiony w rozdziale 2.1.

Rysunek 2.4.3 Schemat architektury sieci neuronowej



Źródło: opracowanie własne

Powyższy schemat przestawia, przedstawia architekturę sieci neuronowej dostosowanej do potrzeb estymacji LGD. Objaśnienie elementów architektury przedstawione w Tabeli 2.4.1:

Tabela 2.4.1 Interpretacja elementów architektury

Dane wejściowego dla PC i RR	Dwie warstwy wejściowa, dwie oddzielne macierzy, składające się z wektorów zmiennych objaśniających dla Probability of Cure(PC) i Recovery Rate (RR)
Sieć PC i Sieć RR	Dwie warstwy ukryte-dwie niezależne sieci neuronowe, służące do obliczenia PC i RR
Wynik PC i Wynik RR	Warstwa wyjściowa dla dwóch sieci.
Wyznaczenie oszacowanego LGD	Warstwa obliczeniowa, wyznacza wartość LGD zgodnie ze wzorem: $LGD = (1 - \mathrm{PC}) * (1 - \mathrm{RR})$
Porównanie z rzeczywistą wartością LGD	Obliczenie średniego błędu kwadratowego, do ustalenia skali błędu popełnionego przez sieć
Backpropagation	Stosowana reguła łańcuchowa dla wykrycia najmniej efektywnych wag i następnie aktualizacja wag

Źródło: opracowanie własne

Ważniejszym elementem sieci neuronowej jest proces uczenia się modelu, realizowany poprzez optymalizacje wartości wag i wartości przesunięcia. W tym celu stosowany jest algorytm propagacji wstecznej (*ang. backpropagation*). Polega on na obliczaniu gradientu średniego błędu kwadratowego względem parametrów modelu, przenosząc informacje o błędzie wstecz: od warstwy wyjściowej do warstwy wejściowej, określając wpływ każdego parametru na błąd predykcji. W ramach niniejszej pracy, przewidywana wartość LGD jest obliczana na podstawie dwóch niezależnych sieci neuronowych, z kolei gradienty są wyznaczane względem wspólnej funkcji straty. Obliczania te realizowane według następującego wzoru[21]:

[21] https://www.geeksforgeeks.org/machine-learning/backpropagation-in-neural-network/ (dostęp dnia 28.06.2025)

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial z} * \frac{\partial z}{\partial w} \tag{20}$$

gdzie:
L- średni błąd kwadratowy

$\hat{y}$ – przewidywana wartość LGD

$\frac{\partial \hat{y}}{\partial z}$ pochodna funkcji aktywacji

Podsumowując, siec neuronowa typu MLP jest skutecznym rozwiązaniem problemu obliczenia strat z tytułu niewypłacalności klienta. Uwzględniając wszystkie jej zalety, należy rozważyć wyżej opisany metod również w praktycznym ujęciu na rzeczywistych danych kredytowych, odpowiednio porównując z bardziej klasycznymi metodami statystycznymi.

Rozdział III. Wstępna analiza danych i badania empiryczne

3.1. Eksploracyjna Analiza Danych

W niniejszej pracy zostały wykorzystane zbiory danych pochodzące z publicznie dostępnego zbioru opublikowanego na platformie Kaggle. Platforma Kaggle jest powszechnie uznawaną analityczną platformą chmurową. Platforma została stworzona w 2010 roku. Od 2017 roku została wykupiona przez Google i stanowi znaczną częścią jego ekosystemu Google. Kaggle pełni funkcje rozbudowanego środowiska, wspierający realizacje projektów analitycznych z różnych obszarów życia społeczno-gospodaczego. Również platforma oferuje również liczną bibliotekę publicznie dostępnych zbiorów danych. Każdy użytkownik ma prawo publikować własne zestawy danych.[22,23]

Zbiory danych wykorzystane w niniejszej pracy pochodzą z materiałów edukacyjnych i zostały przygotowane wyłącznie do celów dydaktycznych. Niemniej jednak dane zostały opracowane na podstawie rzeczywistych schematów klientów sektora bankowego w Indiach.[24] Zestawy danych zawierają informację na temat wniosków kredytowych złożonych przez klientów. Pierwszy zbiór dotyczy danych udzielonych kredytów w szczególności: typ klienta według klasyfikacji wewnętrznej, dane identyfikacyjne i lokalizacyjne klienta, długość relacji z bankiem w miesiącach, wysokość zobowiązania, wartości zabezpieczeń, liczby wcześniejszych zobowiązań, historii spłat oraz data wystąpienia niewypłacalności. Drugi zbiór obejmuje informacje o spłatach zrealizowanych przez klientów, w tym: wysokość wpłat oraz daty ich realizacji. Trzeci zbiór reprezentuję miesięczne salda rachunków klientów. Zbiór kredytowy zawierał 50 000 obserwacji, zbiór ze spłatami kredytów — 626 601 rekordów oraz dane miesięcznych sald obejmowały 400 249 rekordów. Dane pochodzą z okresu od 1 stycznia 2012 do 31 grudnia 2021 dla daty realizacji wypłaty kredytu. W zbiorze również znaleziono obserwacji, posiadające daty niewykonania zobowiązania po okresie 31 grudnia 2024 roku. Wykluczono takie obserwacji z dalszej analizy z powodu zniekształcenie potencjalnych wyników. Również przyjęto, że dla wszystkich analizowanych obserwacji proces windykacji

[22] https://www.geeksforgeeks.org/what-is-kaggle/ (dostęp dnia 05.05.2025)
[23] https://www.kaggle.com/code/arpitasinha12/loan-risk-analysis (dostęp dnia 05.05.2025)
[24] https://www.kaggle.com/datasets/nitishbhardwaj2905/upgrad-bfsi-credit-risk-assignment/data (dostęp dnia 06.05.2025)

został zakończony. Należy zauważyć, że dane obejmują 13-letni okres historyczny. Zbiory zostały połączone za pomocą wspólnej kolumny unikalnego identyfikatora i finalny zestaw, obejmuje 45394 danych klientów.

W tabeli 3.1.1 przedstawione podstawowe statystyki opisowe wybranych zmiennych. Dla zmiennej Kwota kredytu odchylenie standardowe (192 tyś.) przyjmuje wysoką wartość, potwierdzając istnienie znaczących różnic w wysokości udzielanych kredytów. Również rozkład tej zmiennej cechuję się silną asymetrią prawostronną (1.72). To wskazuje na wystąpienie w danych, istotnej liczby wysokich kwot kredytów. Warto zaznaczyć, że dla zmiennych Miesięczna rata i Łączna spłata przed niewypłacalnością, wartość kurtozy wynoszące 11.1 i 7.43 odpowiednio, świadczy o leptokurtycznym rozkład gęstości. Wartości skupiają się wokół średniej wartości, jednak z powodu grubych ogonów dany rozkład zawiera wielu wartości ekstremalnych. Zmienną Wartość zabezpieczenia statystyki opisowe określają jako zmienną o dodatniej skośności(2.72) i istotnie dużym odchyleniem standardowym(90079.21). Innymi słowy, liczna część klientów posiada zabezpieczenie kredytu na stosownie niską kwotę kolei rozkład gęstości zmiennej Oprocentowanie jest zbliżony do rozkładu symetrycznego, gdzie skośność wynosi 0.01, a kurtoza −1.2, nadając rozkładowi płaski kształt. To wynika z braku wyraźnych różnic pomiędzy produktami kredytowymi.

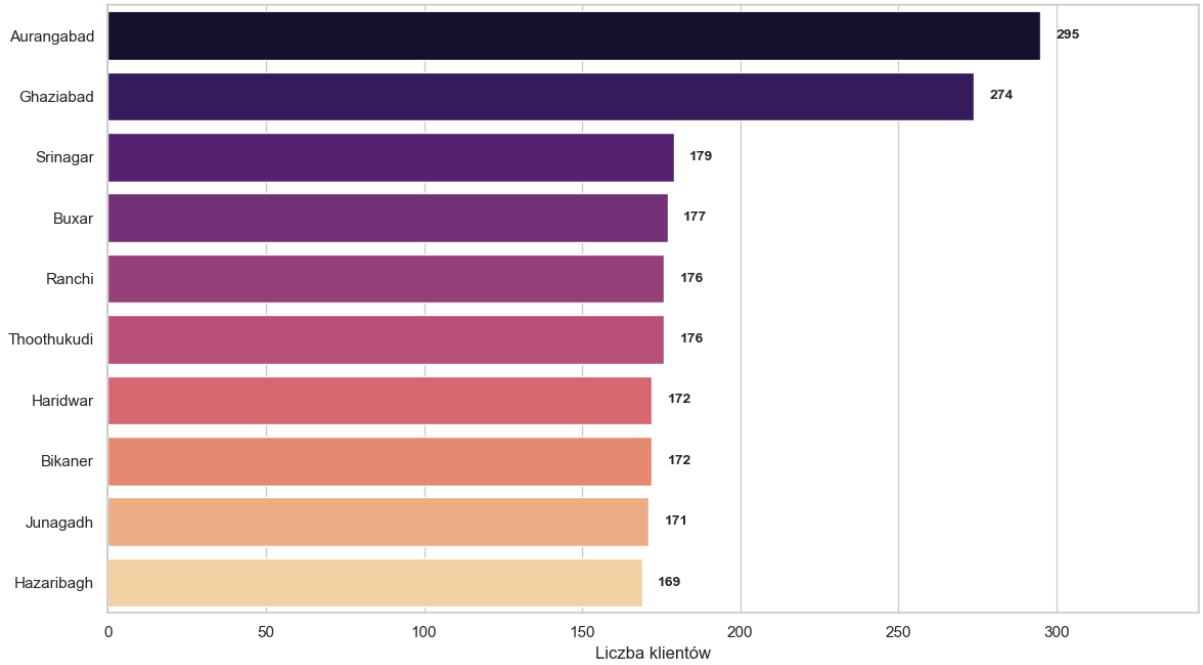
Tabela 3.1.1 Statystyki opisowe zmiennych

Zmienna	**Średnia**	**Mediana**	**Odchylenie Standardowe**	**Wariancja**	**Skośność**	**Kurtoza**
Kwota kredytu	381324.97	191997	503929.94	$2.53*10^{11}$	1.72	1.88
Wartość zabezpieczenia	54291.15	18444.14	90079.21	$8.11*10^{9}$	2.72	7.84
Zwrócone czeki	1.67	1	1.72	2.94	0.9	0.23
Liczba aktywnych kredytów	1.40	1	1.21	1.47	0.54	-0.46
Liczba pominiętych spłat	9.1	7	7.28	53.03	0.88	0.13
Długość relacji z bankiem	84.11	83	43.23	1868.89	0.21	-0.66
Okres kredytowania	2.98	3	1.41	1.99	0.02	-1.3
Oprocentowanie	11.48	11.5	2.02	4.08	0.01	-1.2
Miesięczna rata	16593.37	6550.76	26679.19	$7.11*10^{8}$	3.05	11.1
Łączna spłata przed niewypłacalnością	164600	61766.8	262098.52	$6.86*10^{10}$	2.63	7.43
Uśrednione saldo konta	8157.5	2383.93	16735.86	$2.8*10^{8}$	4.93	34.53

Źródło: opracowanie własne

Na rysunku 3.1.1 pokazuje miasta według największej liczby klientów z nieuregulowanymi zobowiązaniami. Najwięcej przypadków wykryto w mieście Aurangabad (295) oraz Ghaziabad (274). Aurangabad pełni role ważnego ośrodka przemysłowego, z silnie rozwiniętym sektorem motoryzacyjnym, farmaceutycznym i tekstylnym. Natomiast, znaczna część mieszkańców, żyje poniżej granicy ubóstwa. W przypadku Ghaziabad, które jest położone w stanie Uttar Pradesh w północnych Indiach, jest to miasto o szybko rozwijającym sektorze gospodarczym. Jednak wysoki procent ludzi zatrudnionych w szarej strefie oraz brak pewnych źródeł utrzymania, mogą powodować zwiększonego ryzyka niewypłacalności kredytowej. Ważną tendencją jest łączy udział miast Aurangabad i Ghaziabad w danych, który wynosi około 12%. Następne miasta: Srinagar, Buxar, Ranchi czy Thoothukudi, podobną ilość niewypłacalnych klientów. Poniższy wykres może wskazywać na regionalne różnice w polityce kredytowej i poziomie dochodów.

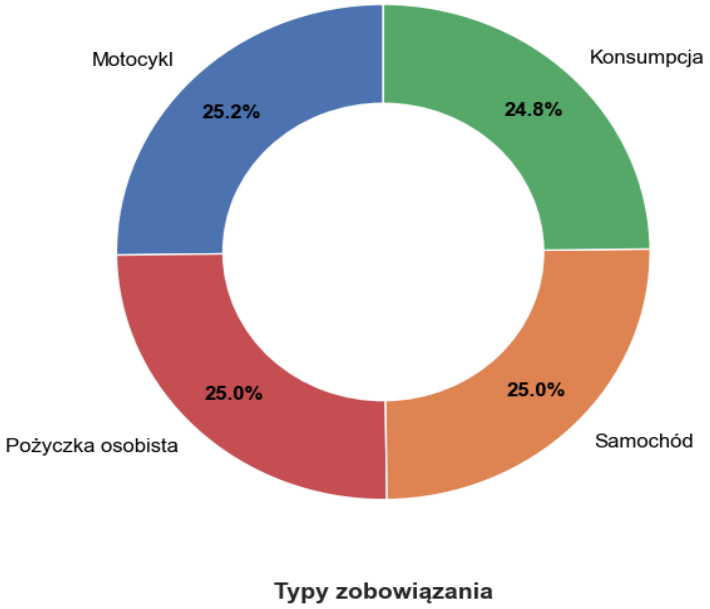
Rysunek 3.1.1 Miasta według największej liczby klientów z zaległością w spłacie zobowiązania



Źródło: opracowanie własne

Rysunek 3.1.2 pierścieniowy, reprezentuję deklarowane cele, dla których udzielane kredyty. Cele stanowią cztery kategorie: motocykl, konsumpcja, pożyczka osobista i samochód. Można zaobserwować, że udziały poszczególnych kategorii zobowiązań są prawie jednakowe. To potwierdza optymalną strukturę ekspozycji kredytowej. Jednak największy udział mają zobowiązania dla zakupu motocykli (25.2%).

Rysunek 3.1.2 Struktura typów kredytów według deklarowanych celów



Źródło: opracowanie własne

Z uwagi na kilku czynników w niniejszej pracy zastosowany jest uproszczony wzór do wyznaczenia LGD. Pierwszym powodem jest brak informacji dotyczących kosztów windykacji zbiorze danych. Drugim powodem jest mniejsza efektywność stosowania klasycznego podejścia IRB, które w danych okolicznościach doprowadzi do przeszacowania ryzyka. Wzór ten, można zapisać następująco:

$$LGD = \frac{loan - (collateral + total\ repayment)}{loan} * 100 \qquad (21)$$

gdzie:
loan-wysokość udzielonego kredytu

collateral- wartość zabezpieczenia

total_repayment- łączna ilość spłat przed stanem niewywiązania z zobowiązań

Następnym krokiem był podział wartości LGD w zależności od kategorii straty. Według Tabeli 3.1.2, najbardziej liczną grupę stanowią klienci, których poziom straty jest bardzo wysoki – 12355 przypadków. Następnie umiarkowane i wysokie straty razem stanowią około 45% całego zbioru (11640 oraz 11517 odpowiednio). Natomiast przypadki, gdzie strata jest niska oraz gdzie suma spłat i wartość zabezpieczeń przekracza kwotę pierwotnie udzielonego kredytu, stanowią mniejszość zbioru.

Tabela 3.1.2 Liczba klientów według kategorii strat LGD na podstawie wzoru 18

Kategoria straty	Logika	Liczba klientów
Bardzo wysokie straty	LGD > 0.6	12 355
Umiarkowane straty	0.2 ≤ LGD ≤ 0.4	11 640
Wysokie straty	0.4 < LGD ≤ 0.6	11 517
Niskie straty	0 < LGD < 0.2	9781

Odzysk większy niż ekspozycja	LGD < 0	101

Źródło: opracowanie własne

Jednocześnie, LGD również zostało wyznaczone na podstawie wzoru (4), który uwzględnia elementy Probability of Cure (PC) oraz Recovery Rate (RR). W kontekście historycznego charakteru danych wzór (4), nadal można uznać za zasadne, gdy głównym celem jest zbudowanie modelu predykcyjnego. Przyjęto, że klient będzie uznany za wyleczonego, jeżeli faktycznie spłacił zobowiązanie. Warto zaznaczyć, że łączna ilość osób wyleczonych w próbie jest ekstremalnie mała i wynosi 101 przypadków.

Tabela 3.1.3 przedstawia podział wartości LGD, wyliczony na podstawie wrozu (4), w zależności od kategorii straty. Można zaobserwować, że w porównaniu do Tabeli 3.1 2, rozkład strat poniesionych przez bank nie uległ znacznej poprawie. Wyniki niemal identyczne. Natomiast kategoria odzysk większy niż ekspozycja migrowała do kategorii niskich strat, z powodu przyjętego założenia wyleczonego.

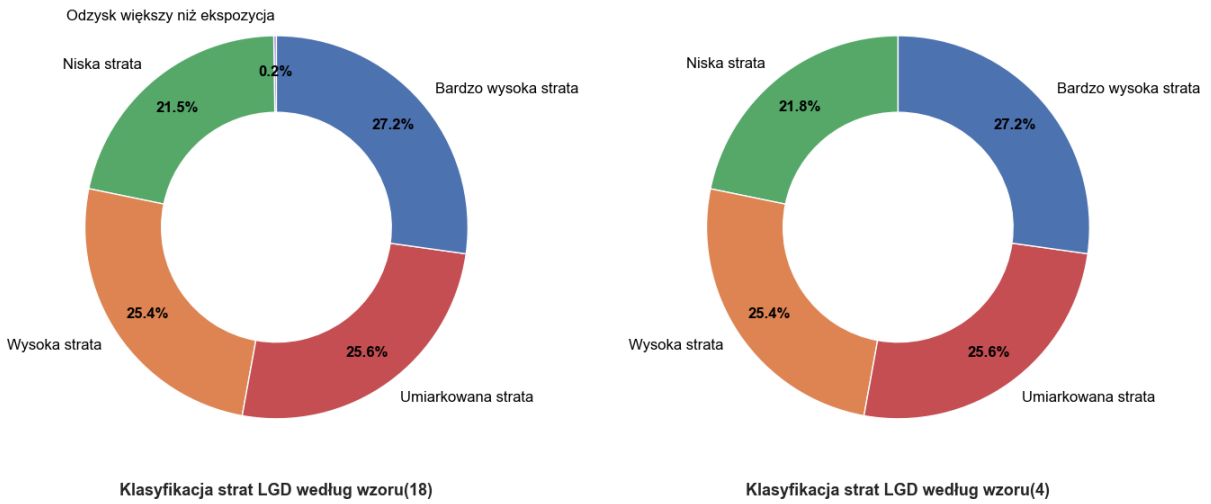
Tabela 3.1.3 Liczba klientów według kategorii strat dla wzoru 4

Kategoria straty	Logika	Liczba klientów
Bardzo wysokie straty	LGD > 0.6	12352
Umiarkowane straty	0.2 ≤ LGD ≤ 0.4	11638
Wysokie straty	0.4 < LGD ≤ 0.6	11517
Niskie straty	0 ≤ LGD < 0.2	9887

Źródło: opracowanie własne

Rysunek 3.1.3 pokazuję procentowy udział poszczególnych kategorii strat dla LGD ze wzorów (18) oraz (4) odpowiednio. Na podstawie rysunku oraz powyższej informacji, można wnioskować, że wyniki wykazują dużą zbieżność

Rysunek 3.1.3 Udział klientów w poszczególnych kategoriach strat

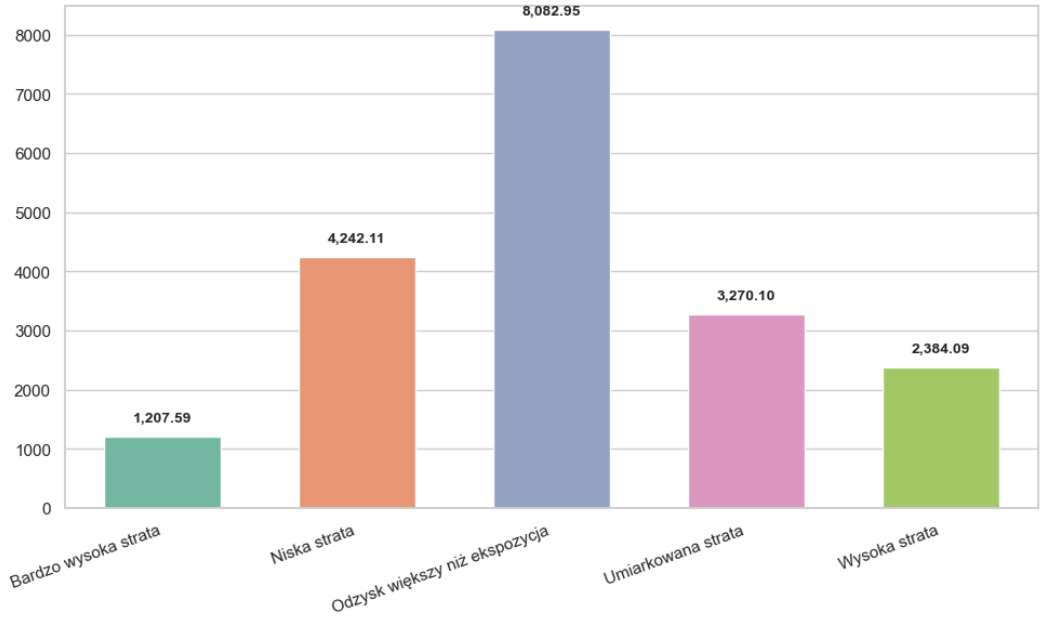


Źródło: opracowanie własne

Rysunek 3.1.4 prezentuje zależność pomiędzy średnią wartością salda konta klienta, a przypisaną klasą straty według wzoru (18). Najwyższa uśredniona wartość salda konta wystąpiło w klasie Odzysk większy niż ekspozycja, gdzie średnie saldo wyniosło 8 082.95 rupii. Z kolei klasy Wysoka strata oraz Bardzo wysoka strata osiągały najniższe średnie saldo

2 384.09 rupii oraz 1 207.95 rupii Poniższy rysunek tylko potwierdza aksjomat, że klienci większymi możliwościami finansowymi generują mniejszą wartość LGD.

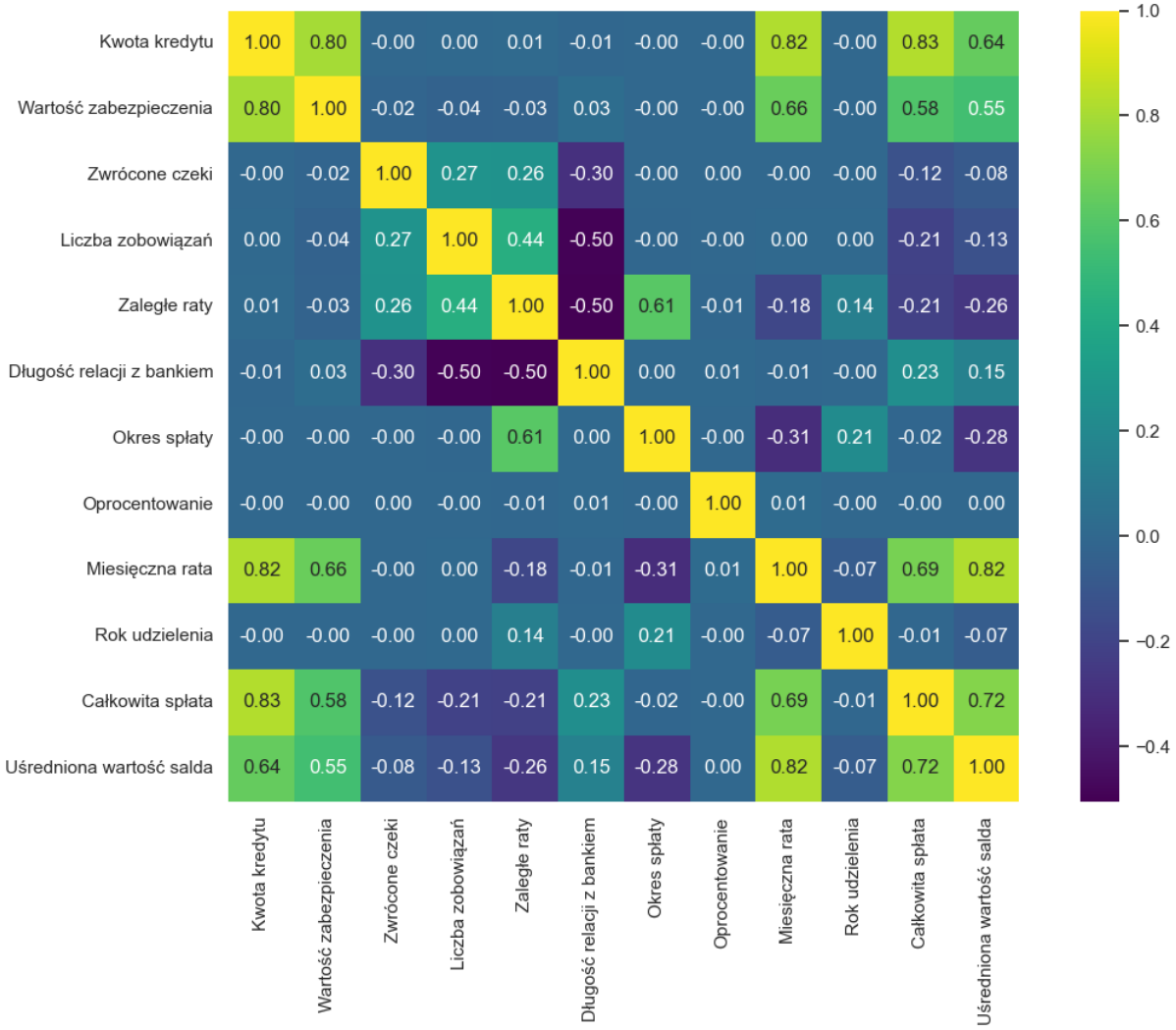
Rysunek 3.1.4 Średnia wartość salda konta, a klasa straty według wzoru 18



Źródło: opracowanie własne

Rysunek 3.1.5 przedstawia macierz korelacji wykonanej w postaci „heatmap", dla wybranych istotnych zmiennych. Bardzo silna dodatnia korelacja (0.80) odnotowana pomiędzy zmiennymi kwotą kredytu a wartością zabezpieczenia. Zaobserwowany wynik wskazuje na stosowanie, przez bank, podobnych kwot zabezpieczeń dla podobnych kwot zobowiązań. Analogicznie wysoka zależność znaleziona również pomiędzy zmiennymi kwota kredytu a całkowita wartość spłaty (0.83). Wyraźna zależność jest związana z faktem, że wyższe wartości kredytu, wymagają wyższych kwot spłat. Następne istotne korelacji występują pomiędzy zmiennej średnia wartość salda a miesięczną ratą (0.82) oraz całkowitą spłatą (0.72). To świadczy o tym, że klienci o wyższym potencjale finansowym, stanowią dla instytucji finansowych niższe ryzyko. Zaobserwowano również zależność pokazującą lojalność wobec banku: zmienna długości relacji z bankiem jest ujemnie skorelowana ze zmienną zaległe raty(–0.50). Tą zależność można interpretować następująco: im większe zaufanie wykazuje bank wobec klienta, tym mniejsze prawdopodobieństwa na nieuregulowanie zobowiązań przez tego klienta. Brak istotnej korelacji cechuję zmienna oprocentowanie z żadną zmienną w wybranym zestawie. To może być spowodowane zbliżoną wartością stop procentowych dla różnych klientów.

Rysunek 3.1.5 Heatmap'a dla wybranego zestawu zmiennych



Źródło: opracowanie własne

Podsumowując, eksploracyjna analiza danych pomogła wykazać dużą ilość istotnych zależności i pokazać prawdziwy charakter danych. Zgodnie z praktyką regulacyjną omawianą w rozdziale 1.1, nie można jednoznacznie określić poziomu wartości LGD uznawanego za bezpieczny. Spowodowane to jest dużą ilością czynników, które wpływają na poziom ryzyka. Niemniej jednak na podstawie przeprowadzonej analizy, możemy stwierdzić, że dla kredytów, zaciąganych w celu zakupu motocykli oraz samochodów wartości LGD na poziomie 0.4 można uznać za bezpieczne dla banku. Natomiast dla konsumpcji oraz pożyczek osobistych wskaźnik LGD 0.5 można klasyfikować jako bezpieczny.

Klienci które nie przekroczyli bezpiecznego progu, charakteryzują się: wysokim średnim saldem konta, historią spłat kredytowych przekraczających 100 000 rupii, poniżej pięciu niespłaconych rat. Dana kategoria klientów często posiada nie więcej niż jedną aktywną pożyczkę. Również wspomniana kategoria klientów posiada długotrwała relacje z bankiem, przekraczająca 100 miesięcy oraz istotną wartość zabezpieczenia o wartości powyżej 100 000 rupii. Opis tego rodzaju klienta może służyć jako rekomendacja bankowi w jego polityce kredytowej.[25]

[25] Nitish Bhardwaj, Shaivi Kharbikar, „BFSI – Credit Risk Assignment", str.1-17

3.2. Estymacja modeli regresji logistycznej

3.2.1 Estymacja regresji logistycznej dla parametru PC

Przed rozpoczęciem modelowania regresji logistycznej dla parametru uzdrowienia klienta, należy uwzględnić niektóre istotne aspekty. W pierwszej kolejności zbudowano dodatkową zmienną: uśredniona wartość stopy odzysku w danym mieście. Głównym powodem utworzenia tej zmiennej było dążenia do skorzystania z informacji o miastach w modelowaniu. Dane zgrupowano po miastach i następnie obliczona średnia wartość stopy odzysku poziomu odzysku dla każdego miasta.

Następnym krokiem był podział zbioru danych na część treningową i testową. Zbiór treningowy, zawierał 36 315 obserwacji. Natomiast zbiór testowy posiadał 9079 rekordów. Podział został dokonany w proporcji 80:20.

Na etapie estymacji przy wyborze zmiennych objaśniających do modelu regresji kierowano się kryteriami poprawności logicznej oraz statystycznej biorąc pod uwagę cechy zmiennych. Przy dokonaniu wyboru zmiennych do modelu wykorzystano metodę Hellwiga[26] oraz metodę stepwise regression.

Metoda Hellwiga, należąca do metod heurystycznych selekcji zmiennych, opiera się na znalezieniu optymalnego podzbioru zmiennych objaśniających wykorzystywanych w modelach predykcyjnych. Główną ideą metody jest obliczenie wskaźnika informacyjnego dla każdego możliwego podzbioru zmiennych i znalezienie kombinacji, która posiada więcej informacji o zmiennej zależnej. Wskaźnik informacyjny opiera na: korelacji pomiędzy poszczególną zmienną objaśniającą a zmienną zależną oraz korelacji pomiędzy zmiennymi objaśniającymi. Metoda Hellwiga dotychczas odgrywa istotną role w doborze zmiennych do modeli statystycznych.

Z drugiej strony metoda stepwise regression, jest również metodą selekcji zmiennych odnosząca do grupy algorytmów deterministycznych. W odróżnieniu od metody Hellwiga, opiera na budowie modelu patrząc na statystyki testowe i wyklucza zmienne nieistotne statystycznie. Ta metoda używa podejść forward oraz backward regression. W algorytmie forward regression, proces selekcji zmiennych sprowadza się do włączenia do modelu pojedynczo istotnych statystycznie zmiennych, zanim zostaną tylko nieistotne statystycznie zmienne. Natomiast algorytm backward regression działa w sposób odwrotny. On buduje model na całym zbiorze zmiennych objaśniających i pojedynczo, usuwa nieistotne statystycznie zmienne z modelu , dopóki nie zostaną wszystkie zmienne istotne.

Za pomocą trzech metod wyróżniono trzy zestawy zmiennych, które są przedstawione w tabeli 3.2.1.1 łącznie ze wskaźnikiem VIF. Zbiory stepwise or forward regression zawierały w sobie zmienne znalezione przez metodę Hellwiga. Z powodu wysokiego wskaźnika VIF, ze zbioru wykluczono zmienne Kwota kredytu oraz Miesięczna rata z modelu z powodu wysokiego VIF.

[26] Metody Ilościowe w Badaniach Ekonomicznych, Tom XII/2, 2011, s. 312–321, „Metoda Hellwiga jako kryterium doboru zmiennych do modeli szeregów czasowych".

Tabela 3.2.1.1 Zmiennej wybrane przez różne metody oraz ich wskaźnik VIF

Zmienna	Metoda Hellwiga	Stepwise regression	Forward regression	VIF
Kwota kredytu	Nie występuję	Występuję	Występuję	5.71
Wartość zabiezpieczenia	Nie występuję	Występuję	Występuję	2.79
Średnia wartość salda klienta	Nie występuję	Występuję	Występuję	3.43
Zaległe raty	Występuję	Występuję	Występuję	3.18
Miesięczna rata	Nie występuję	Występuję	Występuję	7.09
Średnia wartość stopy odzysku w danym mieście	Nie występuję	Nie występuję	Występuję	1.01
Oprocentowanie	Nie występuję	Występuję	Występuję	1.00
Okres spłaty	Występuję	Nie występuję	Nie występuję	2.61
Długość relacji z bankiem	Występuję	Występuję	Występuję	1.83
Liczba zobowiązań	Występuję	Występuję	Występuję	1.55
Zwrócone czeki	Występuję	Występuję	Występuję	1.15

Źródło: opracowanie własne

W tabeli 3.2.1.2 jest przedstawiona pierwsza oraz druga propozycja oszacowanego modelu regresji logistycznej dla predykcji prawdopodobieństwa uzdrowienia klienta. Warto wspomnieć, że dane w próbie treningowej oraz testowej łącznie znajduję się tylko 101 klient, które po okresie niewykonania zaległości uregulowali zobowiązanie i wrócili do regularnych spłat. Z tego względu w danych występuję problem niezbilansowanej próby, który ma istotny wpływ na skuteczność modeli.

W Modelu (1) istotne statystycznie następujące zmienne: zaległe raty, okres spłaty, długość relacji z bankiem, średnia wartość salda klienta oraz liczba zobowiązań. Interpretacje dla poszczególnych zmiennych w Modelu(1) jest następująca: każda dodatkowa zaległa rata zmniejsza szanse klienta na wywiązania z zobowiązania o 11.3% ceteris paribus. Każde dodatkowe zobowiązanie zmniejszę szansę klienta na wyleczenie o 65,6% ceteris paribus. Każdy dodatkowy miesiąc relacji z bankiem zwiększa szansę klienta na wyleczenie o 1,96% ceteris paribus.

Model (2) jest zredukowaną postacią Modelu(1).Obydwa modele zawierają stosunkowo mała wartość pseudo-R^2 wynoszącą 0.3 dla Modelu(1) i 0.2 dla Modelu(2).Mimo braku jednoznacznej interpretacji pseudo- R^2 służy do porównania modeli o podobnych zmiennych objaśniających. Natomiast nie informuję o predykcyjnej sile jak klasyczny R^2 w modelach regresji liniowej.

Następnym krokiem było dokonanie predykcji na podstawie oszacowanych modeli. Oba modele byli uczone na zbiorze treningowym, z kolei predykcji na danych testowych na podstawie oszacowań obu modeli na danych treningowych. Biorąc pod uwagę problem

niezbilansowanych prób, skorzystano z: zasady Cramera oraz ze współczynnika Youdena[27]. Obie metody polegają na znalezienie optymalnego puntku odcjęcia próby. Zasada Cramera zakłada punkt odcięcia dla wartości odpowiadającej odsetkowi obserwacji oznaczonych jako 'wyleczone'". Z kolei współczynnik Youdena zdefiniowany jako:

$$J = Sensitivity + Specificity - 1 \quad (22)$$

Stwierdzono, że współczynnik Youndena jest efektywniejszym od Cramera, kontekście, analizowanych danych. Punkt wyniósł 0.008. Przyjęto założenie, że klasa pozytywna jest oznaczana jako „wyleczeni" klienci, a klasa negatywna „nie wyleczeni" klienci. Na dalszym etapie w oparciu o predykcji obliczono metryki, które znajdują w tabeli 3.2.1.3 ,i pokazują jakość uzyskanych predykcji. Model (1) uzyska największą dokładność równą 94.66%, gdy Model (2) – 92.92%. Wskaźnik recall dla "wyleczonych „klientów w Modelu (1) wynosi 70.83%,i jest on wyższy w porównaniu od Modelu(2) w którym metryka wynosi 58.33%. Model(1) przywiduje poprawnie ponad 70% „wyleczonych", co świadczy o wysokiej skuteczności. Jednakże metryka precision dla "wyleczonych" klientów w obu modelach jest niska (3.43% oraz 2.16% odpowiednio). Niski poziom precyzji można interpretować jako konsekwencja niezbilansowanej próby. Wykresy krzywych ROC przedstawione na rysunku 3.2.1.1 wskazują na lepszą skuteczność Modelu (1) od Modelu(2), bazując na wyniku metryki AUC: 80% dla i 79% odpowiednio. Na bazie rysunku 3.2.1, warto podkreślić, że Model(1) prawidłowo sklasyfikował 17 „wyleczonych" klientów, spośród 24 możliwych, demonstrując skuteczność.

Na podstawie poniższej informacji, możemy stwierdzić, że Model(1) jest jakościowo efektywniejszy od Modelu(2), ze względu na: wyższe wartości metryk oceny modelu, wyższego pseudo-R^2 oraz niższych wartości kryteriów informacyjnych AIC i BIC.

Tabela 3.2.1.2 Oszacowania modeli

Typ	(1)Model	(2)Model
Wartość zabiezpieczenia	0(0)	
Średnia wartość salda klienta	0(0) **	
Średnia wartość stopy odzysku w danym mieście	-5.7655(6.096)	
Zaległe raty	-0.1190 (0.072)**	-0.3958(0.063)***
Okres spłaty	-1.6354 (0.259)***	
Długość relacji z bankiem	0.0195 (0.003)***	0.0132 (0.003)***
Liczba zobowiązań	-1.0646(0.239)***	-0.8660 (0.234)***
Oprocentowanie	0.0426 (0.057)	
Zwrócone czeki	-0.1428 (0.098)	-0.0921 (0.098)
Stała	-1.4903(3.628)	-5.7047(0.525)***
Liczba obserwacji	36315	36315
Pseudo-R^2	0.3	0.2

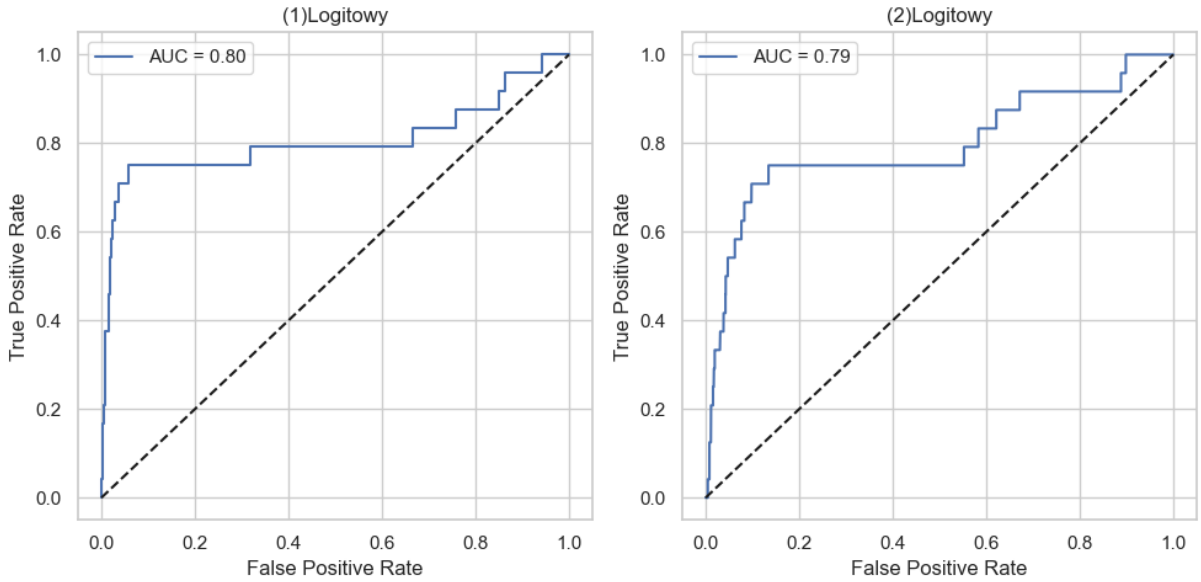
[27] Ośrodek Analiz Statystycznych UMK, dr Joanna Karłowska-Pik, inż. Krzysztof Leki, *Możliwości predykcyjne krzywych ROC*, Uniwersytet Mikołaja Kopernika w Toruniu, str.32-37

AIC	786.60	891.38
BIC	871.60	933.88

Poziomy istotności: *p<0.1;**p<0.05;***p<0.01

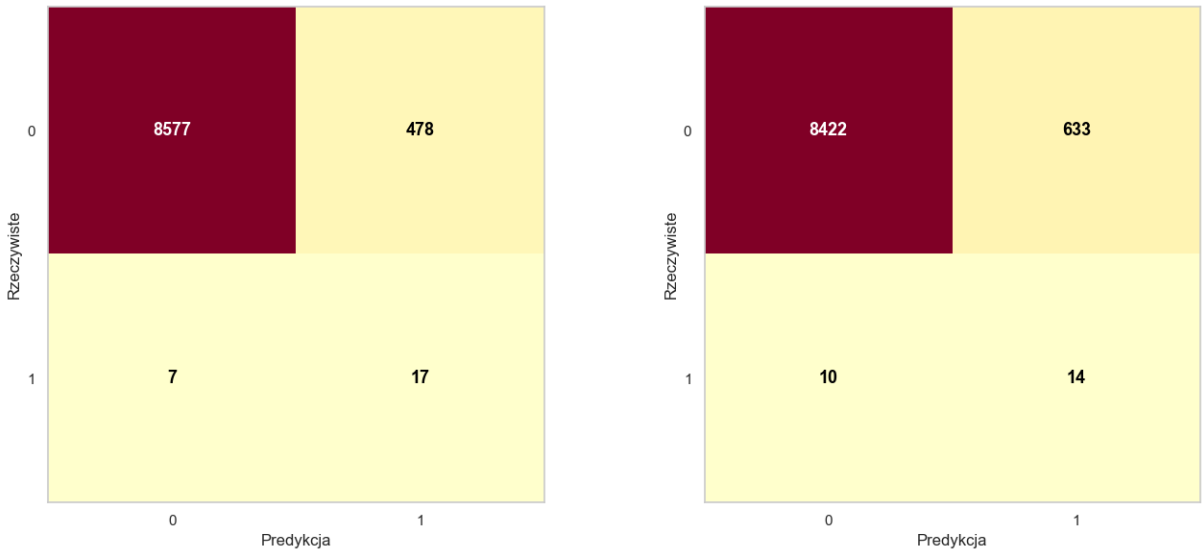
Źródło: opracowanie własne

Rysunek 3.2.1.1 Krzywa ROC dla modeli (1) oraz (2) wraz z miarą AUC



Źródło: opracowanie własne

Rysunek 3.2.1.2 Macierz pomyłek dla modeli (1) i (2)



Źródło: opracowanie własne

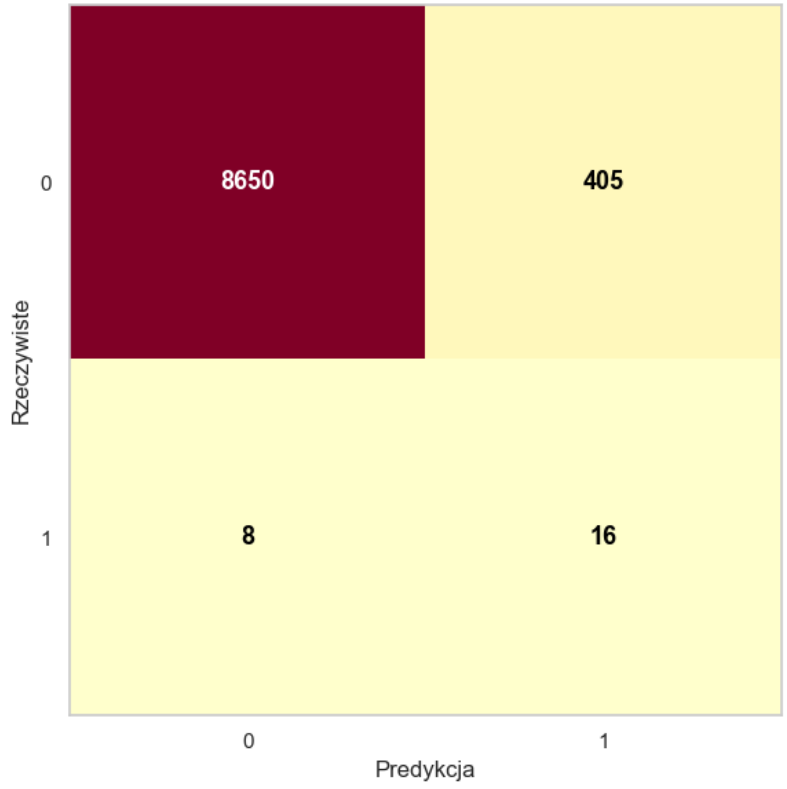
Tabela 3.2.1.3 Metryki dla modeli (1) i (2)

Typ/Metryka	Precision	Recall	Specificity	F1-Score	Accuracy
	Model (1)				
Nie wyleczeni	99.92%	94.72%	94.72%	97.25%	94.66%
Wyleczeni	3.43%	70.83%		6.55%	
	Model (2)				
Nie wyleczeni	99.88%	93.01%	93.01%	96.32%	92.92%
Wyleczeni	2,16%	58.33%		4.17%	

Źródło: opracowanie własne

Również w celu zmniejszenia skali problemu niezbilansowanej próby, skorzystano z algorytmu ADASYN[28] (*ang. Adaptive Synthetic Sampling*). Mechanizm funkcjonowania metody polega na poszukiwaniu obserwacji reprezentujących mniej liczną kategorię oraz wygenerowaniu więcej sztucznych przykładów dla danej kategorii. Zastosowanie algorytmu ADASYN poprawiło wyniki predykcji dla obu klas, które zaprezentowane na rysunku 3.2.1.3.

Rysunek 3.2.1.3 Macierz pomyłek dla Modelu(1)z zastosowaniem ADASYN



Źródło: opracowanie własne

3.2.2 Estymacja regresji logistycznej dla parametru RR

Przed przeprowadzeniem modelowania stopu odzysku, dokonano wyboru zmiennych objaśniających zgodnie z podejściem zastosowanym w rozdziale 3.2.1.Dla wskazanych przez algorytm forward selection zmiennych zbudowano Model(3) oraz Model(4) dla zmiennych określonych przez metodę Hellwiga. Zmienna oraz Miesięczna rata wyłączona z modelu w

[28] https://medium.com/@chirayubaliyan007/adasyn-the-imbalance-data-afd7148f93ef (dostęp dnia 10.05.2025)

związku z wysokim VIF. Zmienna stopa odzysku jest ciągła na całej przestrzeni. 99.5% jej obserwacji mieszczą się w przedziale od 0 do 1. W związku z tym, postanowiono obserwacji poza przedziałem, przypisać do najbliższych maksymalnie dopuszczalnych wartości w przedziale (0,1). I w konsekwencji skorzystano z modeli quasi-binominal dla estymacji zmiennej ciągłej. Rodzina rozkładów quasi-binominal jest używana w sytuacjach, jeśli dane mają charakter binarne, natomiast obserwowana jest większą rozrzutność danych (overdispersion) niż oczekiwano w standardowym modelu binomialnym. Model quasi-binominal wprowadza dodatkowy parametr do modelu, aby uwzględnić tę nadmierną rozrzutność.

Tabela 3.2.2.1 Oszacowania modeli (3) i (4)

Typ	(3)Model	(4)Model
Kwota kredytu	-0.00000022 (0.0000000092)***	
Wartość zabiezpieczenia	0.0000002 (0.000000045)***	
Średnia wartość salda klienta	0.000008 (0.0000002) ***	0.000012 (0.0000002)***
Średnia wartość stopy odzysku w danym mieście	0.831(0.0871)***	
Zaległe raty	-0.1013(0.00063)***	-0.048(0.0005)***
Okres spłaty	0.3301(0.0028)***	
Długość relacji z bankiem	0.0073 (0.000075)***	0.011(0.00009)***
Liczba zobowiązań	-0.1554(0.0024)***	-0.236(0.0028)***
Oprocentowanie	0.00042(0.0012)	
Zwrócone czeki	-0.04477 (0.0015)***	-0.068(0.00174)***
Stała	-0.501 (0.056)***	0.3302(0.012)***
Liczba obserwacji	36315	36315
R^2	0.8392	0.7658
MAE(zbiór testowy)	0.0729	0.0894
RMSE (zbiór testowy)	0.093	0.1123

Poziomy istotności: *p<0.1;**p<0.05;***p<0.01

Źródło: opracowanie własne

W tabeli 3.2.2.1 przedstawione wyniki oszacowanych modeli na danych treningowych. Warto zauważyć, że wszystkie zmienne są istotne statystycznie na poziomie istotności 5% oraz 1% , za wyjątkiem zmiennej Oprocentowanie w Modelu(3). Interpretacje poszczególnych zmiennych są następujące: każda dodatkowa zaległa rata jest związana ze zmniejszeniem stopy odzysku o 9.6 %, ceteris paribus. Każdy dodatkowy rok relacji z bankiem jest związany ze zwiększeniem stopy odzysku o 0.73 %, ceteris paribus. Każdy zwrócony wniosek o kredyt jest związany ze zmniejszeniem stopy odzysku o 4.4% ceteris paribus. Z powodu braku możliwości obliczenia AIC i BIC w modelach rodziny quasi-binominal, stosowano innym zestawem metryk. W Modelu(3) prawie 84% zmienności zmiennej stopa odzysku wyjaśnione poprzez

zmienne objaśniające. Z kolei w Modelu(4) ten wskaźnik jest mniejszy i wynosi 76.58%. W przypadku Modelu(3) średnio myliliśmy około 7.29% , podczas gdy prognozowaliśmy stopę odzysku. Natomiast w Modelu(4) myliliśmy średnio nieznacznie więcej około 8.94%.

Co więcej obliczona została średnia ważona stopa odzysku dla zbiorów testowych. Wyniki obliczeń prezentowane w tabeli 3.2.2.2. Po raz pierwszy Model(3) wykazał nieco gorsze możliwości predykcyjne z wartością delty 0.22% w porównaniu do Modelu(4), dla którego delta wynosi 0.14 %. Należy jednak zauważyć, że obydwa modeli charakteryzowano wysoką trafnością predykcji.

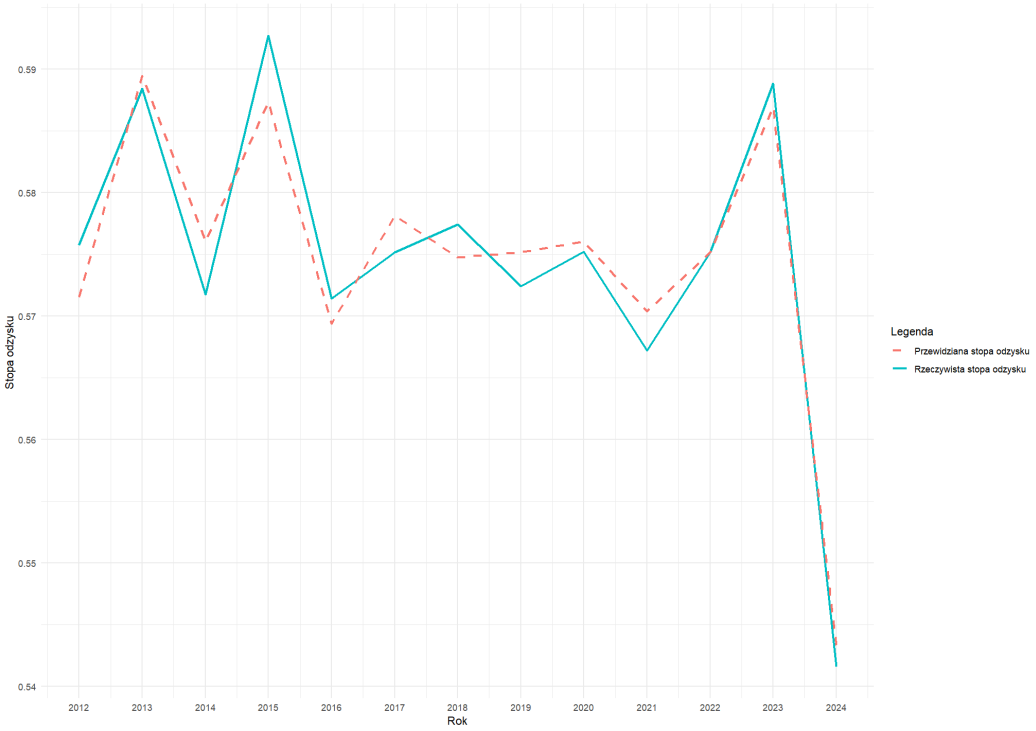
Tabela 3.2.2.2 Rzeczywista i oszacowane stopy odzysku

Model	Rzeczywista średnia ważona stopa odzysku	Oszacowana średnia ważona stopa odzysku	Δ
Model(3)	57.78%	57.56%	0.22%
Model(4)	57.78%	57.92%	0.14%

Źródło: opracowanie własne

Reasumując, Model(3) wykazuję nieznaczną przewagę nad Modelem (4) w przywidywaniu stopy odzysku z powodu: wyższej wartości R^2, mniejszych wielkości metryk MAE,RMSE. Jednocześnie Model(4) posiadał bardziej precyzyjne dopasowanie średniej ważonej stopy odzysku. Na zakończenie warto nawiązać do Rysunku 3.2.2.1, który porównuję średnią szacowaną roczną stopę odzysku i rzeczywistą dla Modelu(3) w latach 2012-2024. Zauważalne jest, pokrywający kształt obu linii, w szczególności dla okresów 2021–2024. To potwierdza powyższe stwierdzenia o trafnym dopasowania modelu do danych testowych.

Rysunek 3.2.2.1 Średnie roczne wartości stopy odzysku rzeczywiste i przewidziane dla Modelu(3), dane testowe



Źródło: opracowanie własne

3.3. Estymacja modeli uczenia maszynowego

3.3.1 Estymacja klasyfikatora XGBoost dla parametru PC

W ramach modelowania prawdopodobieństwa odzyskania należności, za pomocą algorytmu XGBoost, omówiony w rodziale 2.3, zastosowano identyczny zbiór zmiennych objaśniających, na którym utworzono dwa modele regresji logistycznej. Wykorzystany zbiór charakteryzuję się wysokim stopniem niezbalansowania danych, gdzie liczba przypadków wyleczenia stanowi 0.5 % w odniesieniu całego zbioru danych. Zbiór został podzielony próbę treningową i testową w proporcji 80:20. Modele dostosowane są do problemu klasyfikacyjnego z wykorzystaniem logarytmicznej funkcji straty jako funkcji celu.

W ramach znalezienia optymalnego zestawu hiperparametrów dla modelu XGBoost skorzystano z metody optymalizacji bayesowskiej (*ang. Bayesian Optimization*). Algorytm ten, wykazał większą skuteczność w poszukiwaniu optymalnego zestawu zmiennych. Zasadniczą różnicą metody optymalizacji bayesowskiej od metod tradycyjnych w szczególności Grid Search oraz Random Search, jest tworzenie probabilistycznego model funkcji celu. Na podstawie tego modelu wskazuje potencjalne miejsca w przestrzeni parametrów, gdzie kolejne iteracje mogą przynieść najlepsze wyniki, maksymalizując funkcję akwizycji W tabeli 3.3.1.1 przestawione kluczowe hiperparametry w modelu XGBoost, wykorzystane w niniejszej pracy.

Tabela 3.3.1.1 Wybrane hiperparametry w modelu XGBoost

Nazwa parametru	Logika	Uzasadnienie
n_estimators	Liczba drzew	Jeżeli liczba drzew jest mała-niedouczenie,dużo overfitting(rodział 2.1)
gamma	Minimalna poprawa potrzebna do podziału	Zapobieganie tworzeniu nieistotnych podziałów
learning_rate	Waga każdego drzewa	Im mniejszy parametr, tym wolniej model uczy, ale dokładniej
colsample_bytree	Procent zmiennych, które będą użyte do budowy drzew	Ogranicza liczbę zmiennych, które mogą zostać użyte przy tworzeniu każdego drzewa.
red_lambda	L2-regularizacja (Ridge)	Ogranicza złożoność modelu.
red_alpha	L1-regularizacja (Lasso)	Skłania do całkowitego wyzerowania nieistotnych zmiennych z modelu
subsample	Ułamek obserwacji na każde drzewo	Dodaje losowość w modelu. Zapobiega nadmiernemu dopasowaniu modelu.
scale_pos_weight	Waga dla klasy rzadkiej	Ustalana jest odpowiednia waga dla klasy rzadkiej
min_child_weight	Minimalna liczba próbek w liściu	Eliminuje ryzyko tworzenia słabych liści

Źródło: opracowanie własne

Przed uruchomieniem algorytmu poszukującego optymalne wartości parametrów, wyznaczono istotny parametr, który jest brany pod uwagę, gdy występuję klasy rzadka w strukturze zmiennej

objaśnianej. Optymalna wartość scale_pos_weight została wyznaczona, dzieląc liczbę wszystkich rekordów w zbiorze przez liczbę rekordów „wyleczonych". Zakresy oraz optymalne parametry są przedstawione w tabeli 3.3.1.2 dla obu modeli. Warto zwrócić uwagę, że parametr odpowiadający za L2-regularizacje ustalony na wysokim poziomie. To świadczy o mniejszym ryzyku przeuczenia modelu w kontekście klasy nie „wyleczonych". Również learning_rate wykazuje na niskie tempo uczenia. To również wskazuje na dopasowanie parametrów do rzadkiej klasy „wyleczonych".

Tabela 3.3.1.2 Optymalne parametry według metody Bayesian Optimization

Nazwa parametru	Zakres	Optymalny parametr(Model 5)	Optymalny parametr(Model 6)
n_estimators	<100-500>	218	139
gamma	<0-5>	8.06	10
learning_rate	<0.01-0.3>	0.028	0.01
colsample_bytree	<0.3-1>	0.68	0.9
red_lambda	<1-10>	6.47	10
red_alpha	<1-10>	1.418	10
subsample	<0.5-0.8>	0.55	0.5
scale_pos_weight	448	448	470.62
min_child_weight	<5-20>	12.71	17.54

Źródło: opracowanie własne

Następnie optymalne parametry zostały zastosowane dla budowy dwóch modeli o tych samych zmiennych objaśniających. Na danych testowych oceniona jakość ich predykcji wytrenowanego modelu. Na rysunkach 3.3.1.1 oraz 3.3.1.2 pokazany kształt krzywej ROC dla Modelu(5) oraz (6) i również macierzy pomyłek. W pierwszej kolejności, podobnie jak w rozdziale 3.2 przy oszacowaniu modeli regresji logistycznej, zauważalna jest tendencja do przewagi Modelu(5) o większej liczbie parametrów nad Modelem (6). AUC dla Modelu(5) wyniósł 86% natomiast w modelu(6) 81%. Model(6) wykazuję lepszą zdolność do identyfikacji klientów „wyleczonych". Natomiast z macierzy pomyłek wynika, że Model(6) wykazuje lepsze zdolności do identyfikacji klientów „wyleczonych", kosztem gorszej identyfikacji klientów „niewyleczonych". Na podstawie macierzy pomyłek również obliczone metryki: precyzja, czułość, specyficzność i F1-score które przedstawione w tabeli 3.3.1.3. W przypadku Modelu(5) precision dla „wyleczonych" wynosi 3.6%, specificity 41.66% recall 41.67% oraz F1-score równy 6.62%. W przeciwieństwie do tego, w Modelu (6) precision jest na poziomie 1.38%, recall – 58.33%, i wartość F1-score osiągnęła 2.7%. Na podstawie wartości powyższych metryk można stwierdzić że, Model(6) wykazuje lepsze zdolności do identyfikacji klientów „wyleczonych", osiągając wyższe wartości metryki recall. Jednakże Model (5) cechuje się wyższą wartością precision (3.6%). Można to interpretować jako lepszą trafność przewidywań klasy „wyleczonych" w Modelu(5) niż w Modelu(6).

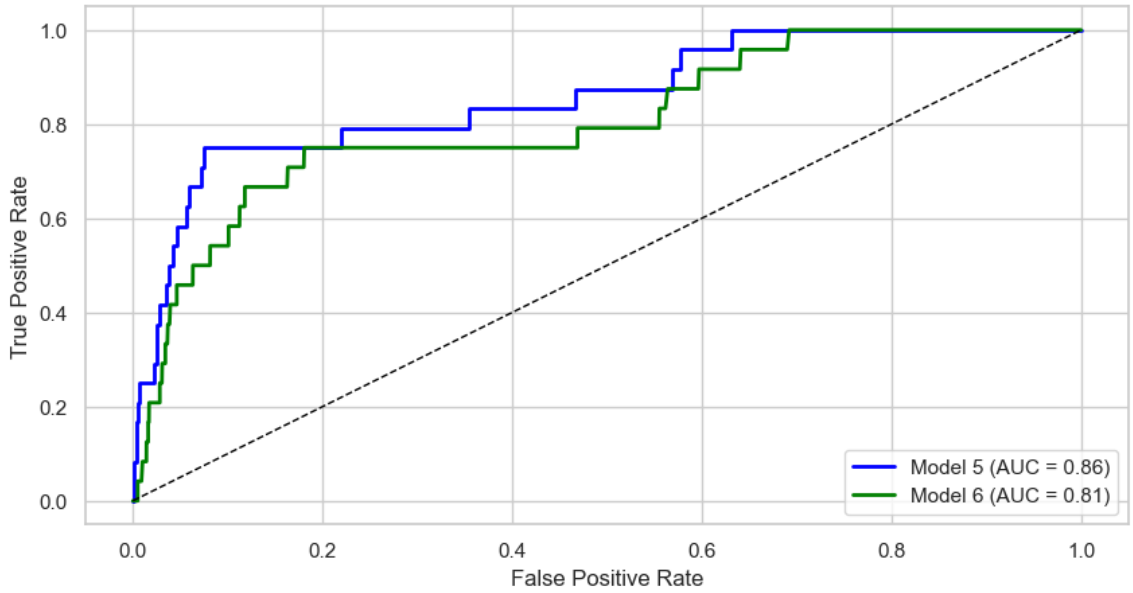
Reasumując, metryki oraz oceny skuteczności wskazują, że Model (5) posiada lepsze zdolności predykcyjne od Modelu (6).

Tabela 3.3.1.3 Metryki dla modeli (5) i (6)

Typ/Metryka	Precision	Recall	Specificity	F1-Score	Accuracy
	Model (5)				
Nie wyleczeni	99.84%	97.04%	99.84%	98.42%	96.89%
Wyleczeni	3.60%	41.67%		6.62%	
	Model (6)				
Nie wyleczeni	99.88%	88.95%	88.96%	94.1%	88.88%
Wyleczeni	1.38%	58.33%		2.69%	

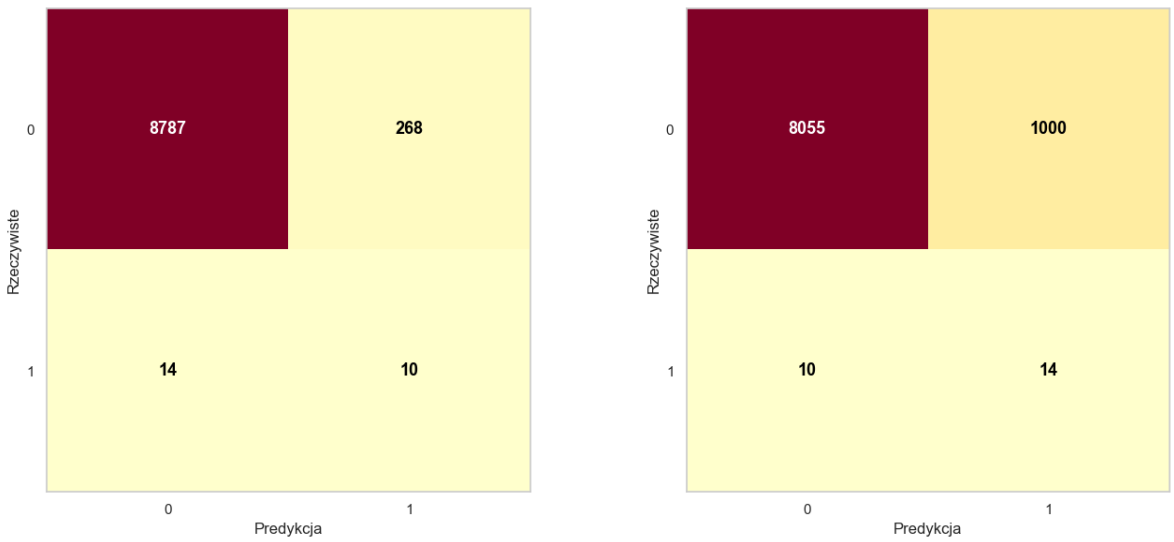
Źródło: opracowanie własne

Rysunek 3.3.1.1 Krzywa ROC dla modeli (5) oraz (6) wraz z miarą AUC



Źródło: opracowanie własne

Rysunek 3.3.1.2 Macierz pomyłek dla modeli (5) i (6)



Źródło: opracowanie własne

3.3.2 Estymacja XGBoost dla parametru RR

W kontekście modelowania wskaźnika odzysku należności za pomocą algorytmu XGBoost wykorzystane zmienne ze zbioru, wybranego w rozdziale 3.2, dla estymacji regresji logistycznej dla parametru RR. Zbiór został podzielony na dwie próbki: treningową i testową w proporcji 80:20. Modele dostosowane są do problemu klasyfikacyjnego z wykorzystaniem pierwiastka z średniego błędu kwadratowego jako funkcji celu. Do zbadania optymalnych hiperparametrów modelu XGBoost, dokładnie jak w zadaniu klasyfikacji dla modelowania PC, skorzystano z algorytmu Bayesian Optimization. Optymalne wartości znalezione przez Bayesian Optimization przestawione w tabeli 3.3.2.

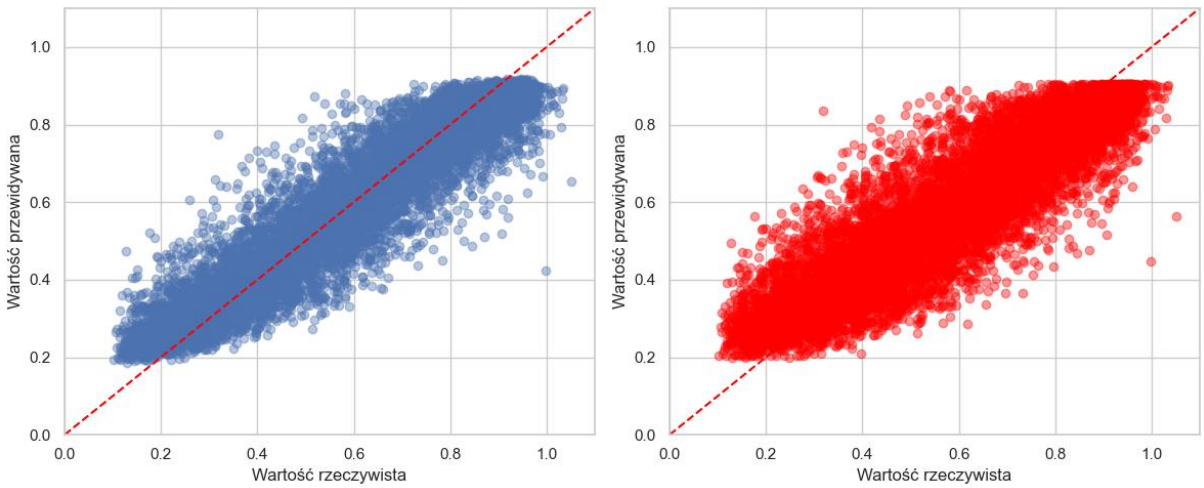
Tabela 3.3.2.1 Optymalne parametry według metody Bayesian Optimization

Nazwa parametru	Zakres	Optymalny parameter(Model 7)	Optymalny parameter(Model 8)
n_estimators	<100-300>	245	276
gamma	<1-10>	1	1
learning_rate	<0.01-0.1>	0.1	0.1
colsample_bytree	<0.5-0.9>	0.9	0.9
red_lambda	<1-10>	2.448	10
red_alpha	<1-10>	1	10
subsample	<0.5-0.8>	0.55	0.8
min_child_weight	<5-20>	5	9.6751

Źródło: opracowanie własne

Kolejnym krokiem zbudowano dwa modele, oceniona jakość ich predykcji na zbiorze testowym. Błąd średniokwadratowy oraz średni błąd absolutny dla Modelu(7) wyniósł 0.13 oraz 0.07 odpowiednio, a dla Modelu(8) 0.13 i 0.08. Model (7) wyjaśnia ponad 77% zmienności odzysku należności. Natomiast Model (8) wyjaśnia 71% zmienności odzysku należności.

Rysunek 3.3.2.1 Wartości rzeczywiste oraz szacowane dla wskaźnika odzysku Model(7) oraz Model(8)



Źródło: opracowanie własne

Rysunek 3.3.2.1 prezentuję, s wykresy rozrzutu wartości przewidywanych względem wartości rzeczywistych dla Modelu(7) oraz Modelu(8). Tego rodzaju wykresy, reprezentują obserwację jako punkt na osi współrzędnych. Przerywana linia czerwona, odnosi do linii idealnego dopasowania. Zjawisko to występuje wtedy, gdy czyli wartość przewidywana jest identyczna wartości rzeczywistej. W przypadku Modelu (7) można rozpoznać duże skupienie punktów

obok czerwonej linii. Kluczowe jest, że nie pojawiają się znaczne odchylenia od linii idealnego dopasowania. Możemy również zobaczyć bardzo małą liczbę wartości odstających na wykresie. Natomiast Model(8) charakteryzuję się większym rozrzutem punktów wokół czerwonej linii. Zasadniczą różnicą pomiędzy Modelem(7) a (8) jest tendencja do większych błędów predykcji, przy większych wartości wskaźnika odzysku. Ta tendencja jest zgodna z wyższymi wartościami błędu średniokwadratowego oraz pierwiastku z błędu średniokwadratowego. A zatem rysunek potwierdza przewagę modelu(7) nad modelem(8).

Tabela 3.3.2.2 Rzeczywista i oszacowane stopy odzysku

Model	Rzeczywista średnia ważona stopa odzysku	Oszacowana średnia ważona stopa odzysku	Δ
Model(7)	57.99%	58.24%	0.25%
Model(8)	57.99%	57.94%	0.05%

Źródło: opracowanie własne

W tabeli 3.3.2.2 reprezentowane wyniki rzeczywistych i oszacowanych stop odzysku dla Modelu(7) i Modelu(8). Warto zaznaczyć, że pomimo osiągnięcia lepszych wyników dla metryk, Model(8) jest precyzyjniejszy w przewidywaniu średniej ważonej stopy odzysku. To implikuje, że Model(8) lepiej dopasowuje do obserwacji o wyższych kwotach zobowiązania.

3.4. Estymacja modelu sieci neuronowych

W ramach modelowania straty, poniesionej w przypadku niewypłacalności z wykorzystaniem sieci neuronowych, który został szczegółowo opisany w rozdziale 2.4, opiera się na identycznym zbiorze zmiennych objaśniających, który został zastosowany w dwóch poprzednich rozdziałach. Zbiór ten wykazuje wysoki stopień niezbalansowania danych, gdzie liczba przypadków wyleczenia stanowi 0.5 % w odniesieniu całego zbioru danych. Zbiór został podzielony próbę treningową i testową w proporcji 80:20. Jako funkcję celu przyjęto średni błąd kwadratowy (*ang. Mean Squared Error*).

Obie niezależne sieci zostały zbudowane zgodnie ze schematem przedstawionym na rysunkach 2.4.1 i 2.4.2. Zarówno w przypadku parametru Probability of Cure, tak i Recovery Rate warstwa wejściowa obejmuje 4 neurony w przypadku modelu skróconego oraz 9 neuronów w modelu pełnym. Pierwsza warstwa ukryta składa się z 64 neuronów, natomiast druga warstwa ukryta zawiera 32 neurony. Warstwa wyjściowa zbudowana jest z pojedynczego neuronu. Niemniej jednak, różnica między sieciami polega na funkcji aktywacji w warstwie wyjściowej: w przypadku Probability of Cure wykorzystano funkcję sigmoid, natomiast dla Recovery Rate zastosowano regresję.

Po uruchomieniu algorytmu, uzyskano wyniki predykcji LGD oraz szczegółowe informacje dotyczące procesu uczenia. Na rysunku 3.4.1 oraz tabeli 3.4.1 przedstawiono zmianę wartości średniego błędu kwadratowego w zależności od liczby iteracji. W celu monitorowania jakości predykcji na danych niewidzianych przez model, algorytm został zaprojektowany w taki sposób, aby po każdej dziesiątej iteracji obliczana była wartość średniego błędu kwadratowego na zbiorze testowym. Pomiar ten wykonywany był przy użyciu aktualnych wartości wag i wyrazów wolnych modelu, wyuczonych do danej iteracji. Rysunek pozwala zaobserwować graficznie systematyczną poprawę jakości predykcji wraz z kolejnymi iteracjami dla obu modeli. Wartość średniego błędu kwadratowego spadła z 0.0303 do 0.0086 dla modelu

zawierającego 9 zmiennych oraz z 0.0273 do 0.0171 dla modelu zawierającego 4 zmienne. Na podstawie danych z tabeli 3.4.1 można stwierdzić, że obu modelach nie występuje zjawisko przeuczenia. Oba modele są również zdolne do dobrej generalizacji, ponieważ ich wartości funkcji straty dla danych uczących i testowych są zbieżne.

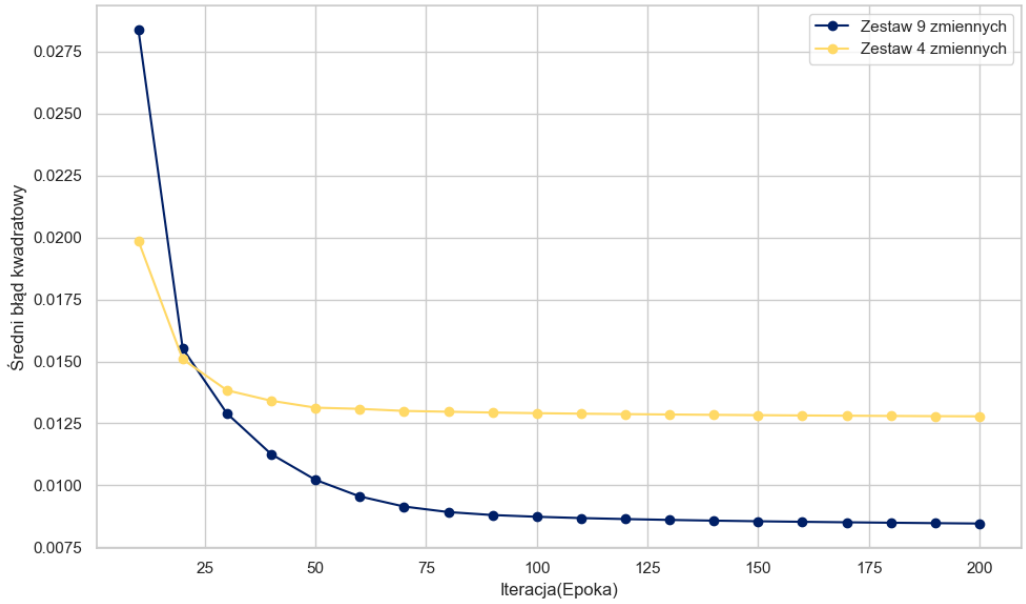
Jednakże, jak i w przypadku poprzednich rozdziałów, model oparty na dziewięciu zmiennych wykazuje lepsze dostosowanie do struktury danych.

Tabela 3.4.1 Zmiana wartości średniego błędu kwadratowego w trakcie uczenia sieci neuronowych dla różnych zestawów zmiennych

Iteracja (epoka)	Średni błąd kwadratowy na zbiorze treningowym(zestaw 9 zmiennych)	Średni błąd kwadratowy na zbiorze testowym(zestaw 9 zmiennych)	Średni błąd kwadratowy na zbiorze treningowym(zestaw 4 zmiennych)	Średni błąd kwadratowy na zbiorze testowym (zestaw 4 zmiennych)
10	0.0303	0.0294	0.0256	0.0273
20	0.0168	0.0194	0.0158	0.0189
30	0.0132	0.0181	0.0146	0.0183
40	0.0116	0.0164	0.0137	0.0176
50	0.0103	0.0153	0.0133	0.0173
60	0.0097	0.0146	0.0132	0.0173
70	0.0092	0.0140	0.0131	0.0172
80	0.0090	0.0136	0.0130	0.0171
90	0.0088	0.0134	0.0130	0.0171
100	0.0087	0.0133	0.0129	0.0171
110	0.0087	0.0133	0.0129	0.0171
120	0.0086	0.0132	0.0129	0.0171

Źródło: opracowanie własne

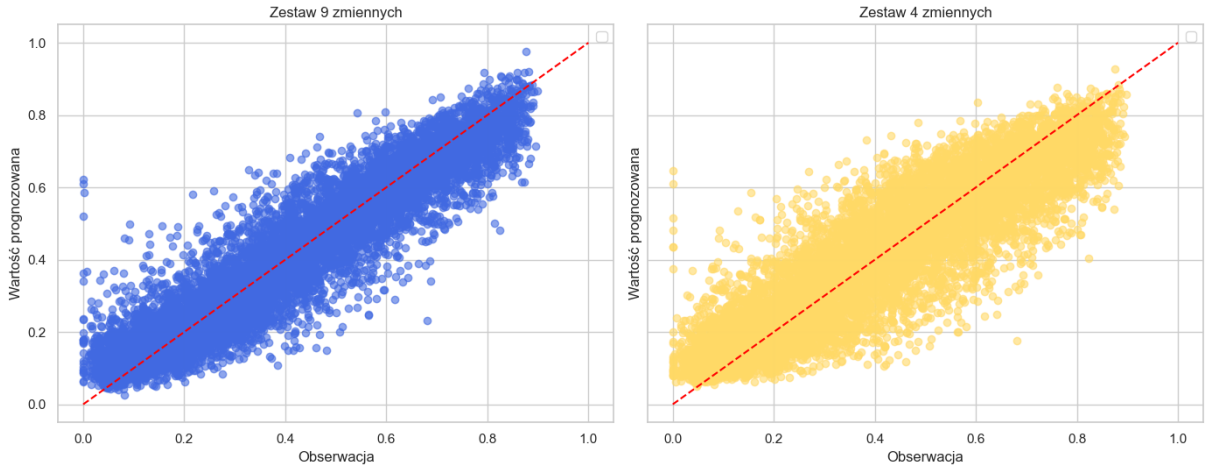
Rysunek 3.4 1 Zmiana wartości średniego błędu kwadratowego podczas uczenia dla dwóch modeli na danych treningowych



Źródło: opracowanie własne

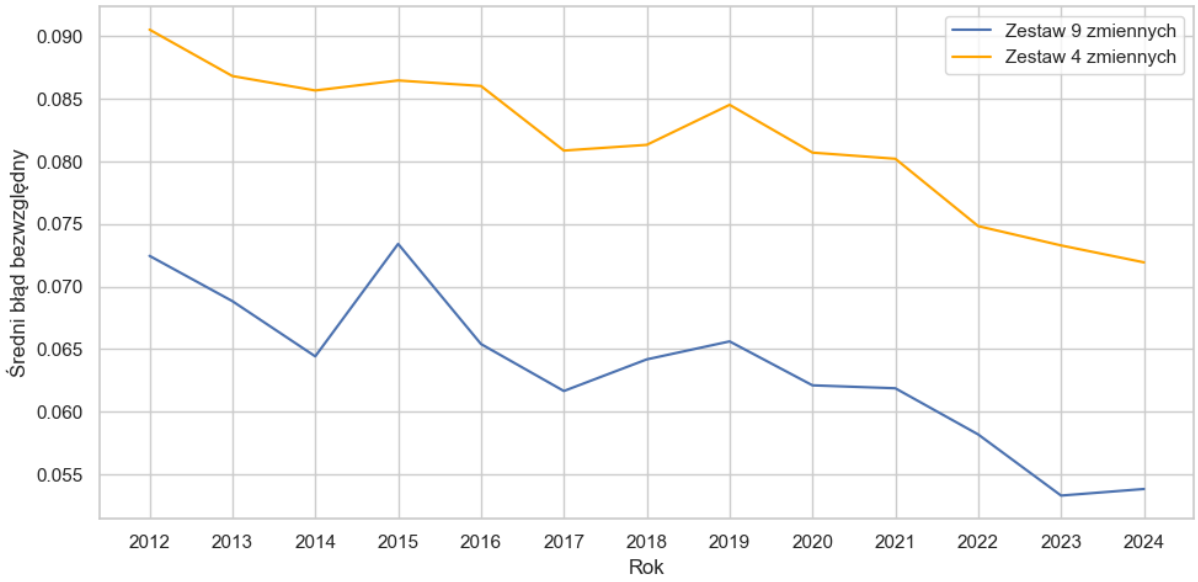
Po przeanalizowaniu przebiegu procesu uczenia modelu, następnym etapem jest ocena dokładności predykcji modelu. W szczególności dokonano analizy porównawczej prognozowanych wartości względem wartości rzeczywistych, przedstawionej na rysunku 3.4.2 oraz analizie średnich błędów bezwzględnych w podziale na lata, zaprezentowanej na rysunku 3.4.3. Na wykresie rozrzutu obserwowana jest większa koncentracja punktów wokół linii idealnego dopasowania w przypadku modelu wytrenowanym na zestawie z 9 zmiennych, niż w przypadku uproszczonego model składającego się z 4 zmiennych. Model oparty na mniejszym zestawie zmiennych wykazuje większe rozproszenie. Wykres liniowy demonstruje średni błąd bezwzględny (*ang. Mean Absolute Error*) przewidywań LGD, obserwowany w latach 2012-2024. Można zauważyć wyraźną przewagę modelu z większym zestawem zmiennych, który osiągał niższe wartości błędu bezwzględnego we wszystkich latach. Warto również zwrócić uwagę na systematyczną tendencję spadkowa błędu bezwzględnego od 2019 roku do 2024. To może być związane ze wzrostem stabilności gospodarczej, która prowadzi do mniejszej zmienności w danych kredytobiorców.

Rysunek 3.4.2 Wykres rozrzutu prognozy od obserwacji wartości LGD



Źródło: opracowanie własne

Rysunek 3.4.3 Średni błąd bezwzględny prognozy LGD w ujęciu rocznym



Źródło: opracowanie własne

3.5. Ocena skuteczności modeli i wybór modelu

Ocena zdolności predykcyjnych modeli regresji logistycznej, algorytmu XGBoost oraz sieci neuronowej została dokonana na podstawie wielu różnych aspektów. Niżej wymieniono kilka ważnych czynników, które można uwzględnić przy porównywaniu trzech rozwiązań:

Tabela 3.5.1 Wybrane kryteria porównania trzech klas modeli

Czynnik/ Typ modelu	Regresja logistyczna	Algorytm XGBoost	Sieć neuronowa
Złożoność modelu	Jak zostało wcześniej ujęte w rozdziale 2.1, regresja logistyczna jest prostym modelem liniowym, który jest skłonny do upraszczania rzeczywistości, a w konsekwencji do niedouczenia modelu.	Modeli typu XGBoost są zaliczane do modeli złożonych. Lepiej identyfikuje złożone zależności między zmiennymi objaśniającymi, a zmienną objaśnianą. Istnieje ryzyko przeuczenia.	Modeli sieci neuronowych należą najbardziej złożonych, spośród wszystkich rozważanych w niniejszej pracy rozwiązań. Podobnie jak i model XGBoost wychwytuje złożnone zależności, natomiast jest narażony na przeuczenie.
Interpretowalność	Regresja logistyczna posiada jednoznaczą interpretację współczynników cech jako wpływu zmiennych objaśniających na logarytm ilorazu szans przynależności do danej klas	Złożoność struktury modelu XGBoost utrudnia jednoznaczną interpretacje.	Modeli sieci neuronowych charakteryzują się niską interpretowalnością, nawet te o mniej złożoną strukturą
Siła predykcyjna	Wyniki prognoz, uzyskane na podstawie oszacowań współczynników regresji logistycznej , mogą być porównane za pomocą następujących metryk: accuracy, recall, F1-score, AUC, precision, macierzy pomyłek itd.	Wyniki prognoz modeli XGBoost mogą być porównane za pomocą następujących metryk: accuracy, recall, F1-score, AUC, precision, macierzy pomyłek, R^2, RMSE, MAE, MSE itd.	Wyniki prognoz modeli sieci neuronowych mogą być porównane za pomocą następujących metryk: accuracy, recall, F1-score, AUC, precision, macierzy pomyłek, RMSE, MAE, MSE itd
Czas uczenia modeli i predykcji	Modeli regresji logistycznej zużywają mniej zasobów ze względu na prostą konstrukcję.	W związku ze swoją złożonością modeli typu XGBoost mogą zużywać więcej zasobów podczas trenowania.	Modeli sieci neuronowych zużywają więcej zasobów podczas trenowania niż nawet modeli XGBoost.
Stabilność predykcji modelu	Regresja logistyczna wyróżnia się wysoką	Modeli XGBoost są mogą być jeszcze	Ze uwagi na losową inicjalizację wag oraz

	stabilnością predykcji. Głównym powodem wysokiej stabilności jest liniowa struktura modelu.	bardziej stabilne pod względem predykcji niż modeli regresji logistycznej. To zależy w dużej mierze od złożoności modelu oraz hiperparametrów.	złożony charakter funkcji aktywacyjnych, predykcje modeli sieci neuronowych, mogą się różnić między uruchomieniami algorytmu, nawet przy tym samym zbiorze danych treningowych.
Brakujące dane	Regresja logistyczna wymaga kompletnych danych w zbiorze treningowym.	Modeli XGBoost samodzielnie identyfikuję sposób traktowania brakujących danych, wybierając optymalną ścieżkę.	Modeli sieci neuronowych są silnie wrażliwe na brakujące dane.

Źródło: opracowanie własne

Biorąc pod uwagę powyższe informacje, możemy ocenić oszacowane modeli w odniesieniu do tych kryteriów.

3.5.1 Porównanie modeli dla Probability of Cure

Pierwszym oszacowanym modelem była regresja logistyczna. Mimo swojej prostoty, regresja logistyczna wyróżnia się wysoką efektywnością pod względem metryk oceny jakości klasyfikacji na zbiorze danych testowym: dla Modelu(1) i (2) AUC wyniósł odpowiednio 80% oraz 79%. Na podstawie wniosków z rozdziału 3.2.1 można stwierdzić, że Model(1) osiąga wyższą dokładność oraz lepszy recall i F1-score dla klasy negatywnej. Model (2) cechuje się niższą precyzją i przewidywaniem klasy pozytywnej. Dla klasy negatywnej oba modele osiągają bardzo wysoką precyzję, recall i F1-score.Macierzy pomyłek, prezentowane na rysunku 3.2.1.2 dla obu modeli, charakteryzują się wysoką skutecznością w identyfikacji przypadków klasy negatywnej. Model(1) poprawnie sklasyfikował 8577 przypadków klasy negatywnej, Model(2) - 8422. W przypadku wykrywana klasy pozytywnej, oba modele osiągają niską skuteczność. Model(1) poprawnie sklasyfikował 17 przypadków, popełniając przy tym 7 błędów. Model(2) zidentyfikował 14 przypadków klasy pozytywnej, przy 10 błędach klasyfikacyjnych tego typu. W rozpatrywanej sytuacji niska skuteczność wykrywania klasy pozytywnej wynika z niezbilansowanej próby.

Podsumowując, mimo że oba modele osiągają wysoką dokładność dla klasy negatywnej, ich skuteczność w wykrywaniu klasy pozytywnej jest ograniczona z powodu niezbilansowanego zbioru danych. Zastosowanie metody ADASYN poprawiło wyniki tylko dla klasy negatywnej, jednocześnie pogorszyła wyniki dla klasy pozytywnej. Jednoznaczne lepsze wyniki osiągnął Model(1), który zawierał 9 zmiennych objaśniających.

Modele oszacowane za pomocą algorytmu XGBoost odznaczają się lepszymi wynikami. Konkluzje przedstawione w rozdziale 3.3.1 można podsumować następująco:

- Model (5) osiągał lepsze wyniki klasyfikacji oraz większą wartość AUC (86%) względem Modelu (6) (81%)
- Zaletą Modelu(5) są wyższe wartości metryk precision oraz F1-score dla klasy pozytywnej.
- Model (6) jedynie charakteryzował się wyższą czułością.

Podczas podjęcia decyzji należy uwzględnić również kontekst biznesowy, nie tylko wartości metryk predykcyjnych. W przypadku modelowania Probability of Cure, istotna jest predykcja ilości klientów, którzy z potencjalnie nie spłacą zobowiązania, niż tych, którzy jego spłacą w całości. Jest to związane z tym to, że instytucji finansowej zależy na odpowiednim oszacowaniu wysokości rezerw na wypadek niewywiązania klientów z zobowiązania. Przeprowadzona analiza wskazuje, że Model (5) oparty na algorytmie XGBoost, jest lepszy w zakresie predykcji klasy negatywnej. Natomiast Model(1) regresji logistycznej jest lepszy w zakresie predykcji klasy pozytywnej. Dlatego ostateczną rekomendacją, w kontekście tego zbioru danych, jest model oparty na algorytmie XGBoost.

3.5.2 Porównanie modeli dla Recovery Rate

Ze względu na przewidywaną, jeszcze podczas planowania, trudnością z porównaniem oszacowań modeli regresji logistycznej oraz modeli opartych na algorytmie XGBoost dla wskaźnika Recovery Rate, zastosowano alternatywne podejście. Rezygnując z klasycznego przekształcenia zmiennej ciągłej do postaci binarnej , skorzystano model quasi-binomial. Ten model umożliwia zachowanie założeń regresji logistycznej przy właściwościach regresji ciągłej. Należy jednak zauważyć, że porównanie modeli typu quasi-binomial z modelami klasycznej regresji liniowej należy je rozpatrywać ostrożnie, ponieważ oba modele mogą zachowywać różnie przy różnych zbiorach danych.

Na podstawie konkluzji z rozdziałów 3.2.2 oraz 3.3.2 stworzono tabelę 3.5.2.1, która przedstawia wartości metryk, służących do oceny jakości predykcyjnych. Na podstawie analizy można jednoznacznie stwierdzić, że Model(3) uzyskał najlepsze wartości metryk spośród porównywanych modeli. Jest to model typu quasi-binominal stworzony na podstawie 9 zmiennych, który wyjaśnia około 84% obserwowanej zmienności story odzysku. Należy również zauważyć, że Model(3) cechuje się jednym z najniższych wartości średniego bezwzględnego błędu(0.0729) oraz najniższym średnim błędem kwadratowych(0.093). Model(8) wykazuje najniższą wartość różnicy pomiędzy rzeczywistą stopą odzysku, a prognozowaną(0.05%).Pozostałe modele wykazały wyniki o średnim poziomie trafności.

Tabela 3.5.2.1 Metryki oceny jakości predykcyjnych dla modeli stopy odzysku

Model/ Metryka	R^2	MAE	RMSE	Delta stopy odzysku
Model(3)	83.92%	0.0729	0.093	0.22%
Model(4)	76.58%	0.0894	0.1123	0.14%
Model(7)	77%	0.067	0.1164	0.25%
Model(8)	71%	0.0843	0.1316	0.05%

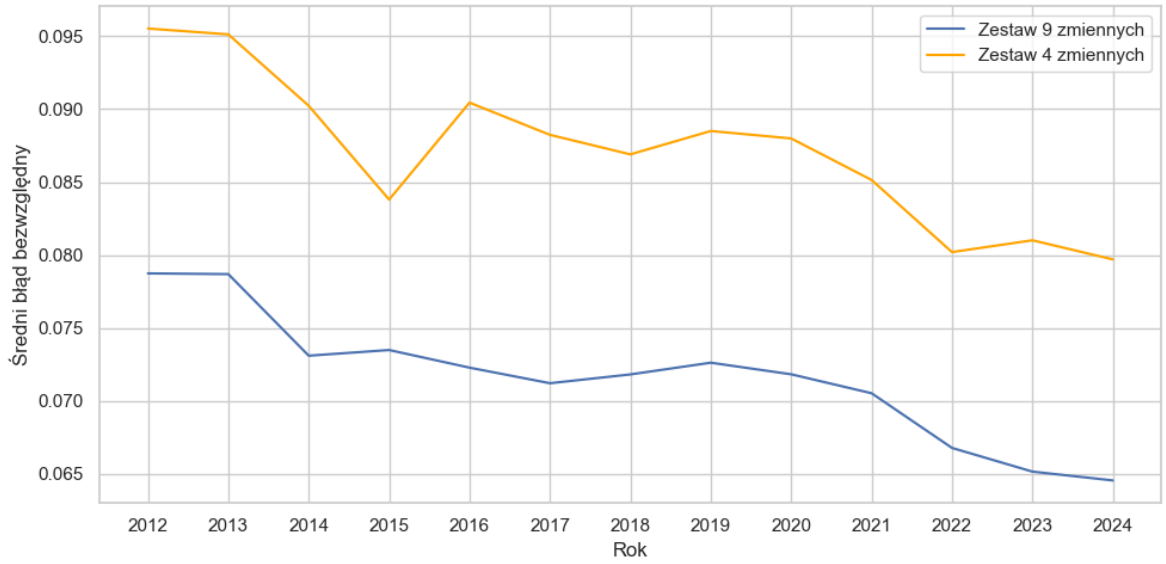
Źródło: opracowanie własne

Podsumowując powyższe rozważania, można wnioskować, że Model(3) osiąga najlepsze wyniki pod względem zarówno jakości predykcji jak i stabilności model, mimo tego, że najniższa delta stopy odzysku jest uzyskana poprzez Model(8). Niska delta sugeruje, że Model(8) dobrze trafia w średnią wartość, natomiast nie zawiera informacji na temat przewidywań konkretnych obserwacji. Również Model(8) potrzebuje więcej zasobów do trenowania. W konsekwencji, z punktu widzenia tego zestawu danych, rekomendowany jest model quasi-binominal.

3.5.3 Porównanie modeli dla Loss Given Default

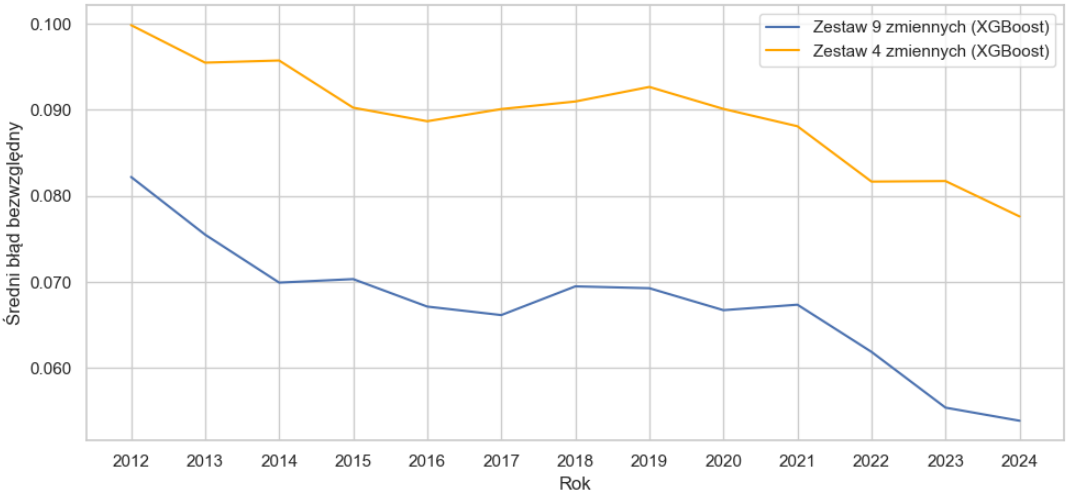
W dwóch poprzednich rozdziałach przeanalizowano poszczególne elementy składowe parametru LGD. Natomiast niniejszy rozdział poświęcony jest przedstawieniu wyników końcowych LGD, uzyskanych, na podstawie podstawienia do wzoru (4) wcześniej oszacowanych komponentów: Recovery Rate oraz Probability of Cure.
Ze względu na odporność na wpływ wartości odstających oraz łatwością interpretacji skorzystano ze średniego błędu bezwzględnego, jako metryki oceny skuteczności modeli. Na rysunkach 3.5.3.1 oraz 3.5.3.2 przedstawione są średnie błędy bezwzględne w podziale na lata dla modeli oszacowanych przez regresje logistyczną oraz modeli XGBoost. Dla sieci neuronowej średni błąd bezwzględny jest prezentowany na rysunku 3.4.3. Na każdym wykresie można zauważyć trend spadkowy wartości średniego błędu bezwzględnego w okresie 2019-2024. Większa dostępność oraz jakość danych, może być powodem tego spadku. Analiza rysunków wyraźnie potwierdza, że najwyższą rzetelnością prognoz charakteryzuje się model sieci neuronowej oparty na zestawie dziewięciu zmiennych objaśniających. Model ten osiągał najniższe wartości MAE w każdym okresie. Drugą pozycję zajął model XGBoost, który mimo uproszczonej struktury w porównaniu do modelu sieci neuronowej nie pozostał daleko w tle. Również najmniej precyzyjne wyniki uzyskano w przypadku regresji logistycznej, co jest zgodne z wcześniejszymi założeniami.

Rysunek 3.5.3.1 Średni błąd bezwzględny prognozy LGD w ujęciu rocznym dla regresji logistycznej



Źródło: opracowanie własne

Rysunek 3.5.3.2 Średni błąd bezwzględny prognozy LGD w ujęciu rocznym dla modeli XGBoost



Źródło: opracowanie własne

Wnioski

Analiza porównawcza regresji logistycznej oraz modelu opartym na algorytmie XGBoost w kontekście modelowania Probability of Cure i Recovery Rate jako oddzielnych składników LGD potwierdziło występowanie istotnych różnic pomiędzy modelami. Wnioski z rozdziału 3.5 wskazują, że algorytm XGBoost okazał się najlepiej dopasowany w kontekście strategii biznesowej banku dotyczącej modelowania Probability of Cure. Strategia biznesowa banku głównie oparta jest nie tylko na uzyskaniu jak najlepszego wyniku predykcji, ale również na zarządzaniu ryzykiem kredytowym. W kontekście zarządzania ryzykiem model oparty na algorytmie XGBoost okazał najlepszy, wykazując większą wrażliwość na obserwacje klasyfikowane jako „nie wyleczeni". Prognoza ilości „nie wyleczonych" obserwacji uznawana za najwyższy priorytet w kontekście modeli LGD, ponieważ przygotowuje instytucję finansową na potencjalne negatywne scenariusze, minimalizując straty pochodzące z braku spłat kredytu. Natomiast w przypadku Recovery rate największe dopasowanie do danych i najwyższą jakość prognozy wykazał model quasi-binominal. Obliczono również wartości LGD na podstawie uzyskanych poszczególnych elementów omówiony w rozdziale 3.5.3. Na podstawie analizowanych danych można wnioskować, że w zakresie predykcji LGD, model sieci neuronowej okazał się najbardziej efektywny. Zbudowane modele na podstawie regresji logistycznej, mimo faktu, że najlepiej przywidują liczbę „wyleczonych" klientów, wymagałby znacznej optymalizacji, aby dorównać do wyników algorytmu XGBoost albo sieci neuronowej. Możliwości rozwoju dla modelu regresji logistycznej dla Probability of Cure obejmują:

- Ponieważ pierwszym naturalnym ograniczeniem przy modelowaniu była niezbilansowana próba, z tego względu, zastosowanie algorytmu ADASYN, który generuje syntetyczne przykłady klasy mniejszościowej, skupiając na klasie problematycznej byłoby zasadne.
- Przeprowadzenie tuningu hiperparametrów regresji logistycznej oraz algorytmu ADASYN.

Niemniej jednak, zgodnie z obowiązującymi wytycznymi nadzorczymi, instytucje finansowe zobowiązane są do stosowania modeli, które charakteryzują się wysokim poziomem interpretowalności i transparentności. Współczesne wytyczne nadzorcze nie akceptują modeli typu "black-box" o ograniczoną interpretowalności. W efekcie, modeli zbudowane w oparciu o algorytm XGBoost oraz modeli sieci neuronowych nie mogą być stosowane ze względu na procedurę zarządzania ryzykiem kredytowym. Jednakże, obecnie część organów regulacyjnych rozważa modyfikację wytycznych, doceniając wartość praktyczną modeli o złożonej strukturze.[29]

Warto podkreślić na zakończenie ograniczenia, związane ze zbiorem danych, na którym budowane modele oraz na którym sformułowane powyższe wnioski. Poniżej przedstawiono kilka z nich:

- Silne niezbilansowanie klas pomiędzy liczbą obserwacji należących do klasy pozytywnej a klasy negatywnej
- Ograniczona informacyjność części zmiennych kategorialnych
- Brakiem istotnych informacji na temat procesu windykacji oraz zewnętrznych zmiennych makroekonomicznych
- Brak danych o scoringu kredytowym klientów

[29] European Banking Authority, Machine Learning for IRB Models: Follow-up Report from the Consultation on the Discussion Paper on Machine Learning for IRB Models, EBA/REP/2023/28, August 2023.

Bibliografia

1. Wojciech Kuryłek, Modelowanie ryzyka portfela kredytowego. Część I, str. 1
2. https://www.knf.gov.pl/dla_rynku/pakiet_crd4/historia_zalozenia#:~:text=%C5%B9r%C3%B3d%C5%82o%20regulacji%20bazylejskich%20si%C4%99ga%201974,marki%20niemieckiej%20i%20dolar%C3%B3w%20ameryka%C5%84skich (dostęp 01.04.2025)
3. Wojciech Kuryłek, Modelowanie ryzyka portfela kredytowego. Część I, str. 6
4. Komisja Nadzoru Finansowego, *BION w bankach – mapa klas ryzyka i ich definicje*, KNF. str. 1
5. Basel Committee on Banking Supervision, 2005, s. 8
6. Basel Committee on Banking Supervision. (2017). *Basel III: Finalising post-crisis reforms*. Bank for International Settlements, s. 56.
7. Komisja Nadzoru Finansowego, *Zarządzanie ryzykiem modeli w zakresie działalności podmiotów sektora bankowego ze szczególnym uwzględnieniem modeli ryzyka kredytowego – podstawowe zagadnienia*, Paweł Grodź, Bartosz Lewandowski, Kamil Simka, Maja Tuszyńska, Warszawa 2024. s. 53
8. Wiszniowski E., *Model szacowania utraty wartości instrumentów finansowych w założeniach MSSF 9 – rachunkowość czy inżynieria finansowa?*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 313, 2013, s. 170–188.
9. European Banking Authority (EBA), *Guidelines on the application of the definition of default under Article 178 of Regulation (EU) No 575/2013*, Final Report, EBA/GL/2016/07.
10. https://www.openriskmanual.org/wiki/Realised_LGD (dostęp 06.04.2025)
11. European Banking Authority, *Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures (EBA/GL/2017/16)*, 2017, s. 71–72
12. Statistical Learning, Trevor Hastie Robert Tibshirani, Springer,2021, str 56-57
13. Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. Annals of Statistics, 29(5), 1189–1232.
14. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
15. https://www.geeksforgeeks.org/xgboost/ (dostęp dnia 25.04.2025)
16. https://medium.com/@prathameshsonawane/xgboost-how-does-this-work-e1cae7c5b6cb (dostęp dnia 26.04.2025)
17. https://fineproxy.org/pl/wiki/regularization-l1-l2/ (dostęp dnia 26.04.2025)
18. Zhu Y., Huang Z., Zhang X., *RS-Boosting and RS-MultiBoosting: Hybrid Ensemble Machine Learning Models for SME Credit Risk Assessment in Chinese Supply Chain Finance*, Forecasting, 2022, Vol. 4, pp. 188–190.
19. Komisja Nadzoru Finansowego, Zarządzanie ryzykiem modeli w zakresie działalności podmiotów sektora bankowego ze szczególnym uwzględnieniem modeli ryzyka kredytowego – podstawowe zagadnienia, Paweł Grodź, Bartosz Lewandowski, Kamil Simka, Maja Tuszyńska, Warszawa 2024. s. 65-66

20. https://www.geeksforgeeks.org/machine-learning/backpropagation-in-neural-network/ (dostęp dnia 28.06.2025)
21. https://www.geeksforgeeks.org/what-is-kaggle/ (dostęp dnia 05.05.2025)
22. https://www.kaggle.com/code/arpitasinha12/loan-risk-analysis (dostęp dnia 05.05.2025)
23. https://www.kaggle.com/datasets/nitishbhardwaj2905/upgrad-bfsi-credit-risk-assignment/data (dostęp dnia 06.05.2025)
24. Nitish Bhardwaj, Shaivi Kharbikar, „BFSI – Credit Risk Assignment", str.1-17
25. Metody Ilościowe w Badaniach Ekonomicznych, Tom XII/2, 2011, s. 312–321, „Metoda Hellwiga jako kryterium doboru zmiennych do modeli szeregów czasowych".
26. Ośrodek Analiz Statystycznych UMK, dr Joanna Karłowska-Pik, inż. Krzysztof Leki, *Możliwości predykcyjne krzywych ROC*, Uniwersytet Mikołaja Kopernika w Toruniu, str.32-37
27. European Banking Authority, Machine Learning for IRB Models: Follow-up Report from the Consultation on the Discussion Paper on Machine Learning for IRB Models, EBA/REP/2023/28, August 2023.
28. https://medium.com/@chirayubaliyan007/adasyn-the-imbalance-data-afd7148f93ef (dostęp dnia 10.05.2025)
29. M. Gruszczyński, *Empiryczne finanse przedsiębiorstw. Mikroekonometia finansowa*, Warszawa: Difin, 2012. str. 1-175
30. J. Kłopotowski, J. Winnicka, *Równania różniczkowe zwyczajne. Teoria i zadania*, str. 1-18
31. Red. nauk.: M. Gruszczyński i in., *Mikroekonometria. Modele i metody analizy danych indywidualnych*, Warszawa: Wolters Kluwer, 2012, 1-195
32. https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstneunet.html (dostęp dnia 20.07.2025)

Spis rysunków

Spis tabel

Streszczenie

Niniejsza praca rozpatruje różne podejścia do modelowania komponentów wskaźnika straty w wypadku niewypłacalności (LGD) dla zobowiązań kredytowych. Brane pod uwagę trzy typy modeli: regresja logistyczna, oparte na algorytmie XGBoost oraz sieci neuronowej. Badanie koncentruje się na dwóch głównych komponentach LGD: prawdopodobieństwo powrotu do stanu wypłacalności (Probability of Cure) oraz stopa odzysku (Recovery Rate). W pracy zostało omówione znaczenie parametru LGD w kontekście ryzyka kredytowego oraz teoretyczne podstawy modeli predykcyjnych Przedstawiono również szczególna analiza zbioru danych, empiryczne zastosowanie modeli, metody oceny ich efektywności i wyciągnięte wnioski na podstawie empirycznych wyników. Regresja logistyczna i XGBoost zostały zastosowane do predykcji PC i RR. Ich wyniki zostały porównane na podstawie metryk oceny predykcji, takich jak AUC, specyficzność, precyzja dokładność, F1-score i czułość. Również wszystkie trzy modeli zastosowano do predykcji LGD. Rezultaty badania dowodzą, że model sieci neuronowej oparty na 9 zmiennych objaśniających osiągnął lepsze wyniki dla modelowania LGD. Z kolei model zbudowany z 9 zmiennych objaśniających, oparty na algorytmie XGBoost, był bardziej efektywny dla modelowania PC, cechując się większą wrażliwością na przypadki „nie-wyleczonych" klientów. Równocześnie model quasi-binomial, składający z 9 zmiennych objaśniających, pochodzący z rodziny modeli regresji logistycznej, okazał się najskuteczniejszy dla modelowania RR. Rezultaty analizy wskazują na potrzebę dalszych badań nad optymalizacją modelu regresji logistycznej.

Oświadczenie autora o samodzielnym wykonaniu pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami. Oświadczam również, że przedstawiona praca dyplomowa nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni. Oświadczam ponadto, że niniejsza wersja pracy dyplomowej jest identyczna z załączoną wersją elektroniczną.

05.08.2025 Dzmitry Fiodarau