DEBIUTY STUDENCKIE

2023

INFORMATYKA W BIZNESIE

pod redakcją
Heleny Dudycz

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2023

Recenzja naukowa
Marcin Hernes

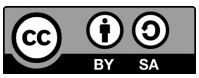
Redakcja wydawnicza
Małgorzata Tadrzak-Mazurek

Korekta
Aleksandra Śliwka

Skład i łamanie
Małgorzata Myszkowska

Projekt okładki
Beata Dębska

Na okładce wykorzystano zdjęcie z zasobów Adobe Stock



ISBN 978-83-67400-80-0 (wersja papierowa)
ISBN 978-83-67400-81-7 (wersja elektroniczna)

DOI: 10.15611/2023.81.7

Druk i oprawa: TOTEM

Andrzej Małowiecki

e-mail: 174415@student.ue.wroc.pl
ORCID: 0009-0000-1264-1436

Uniwersytet Ekonomiczny we Wrocławiu

Metody resamplingu danych w rozwiązaniu problemu nierównowagi danych przy wykrywaniu oszustw związanych z kartami kredytowymi

DOI: 10.15611/2023.81.7.08
JEL Classification: Y80

Streszczenie: Nierówność danych (ang. *data imbalance*) jest jednym z najpowszechniejszych problemów przy zadaniu klasyfikacji. W uczeniu maszynowym próbka danych ma stanowić wiarygodne odzwierciedlenie całości populacji. Jednak równie istotny jest fakt, żeby była zbudowana w sposób, który zapewni modelowi najlepsze warunki w procesie uczenia. Znalezienie równowagi między tymi dwoma aspektami stanowi jedno z wyzwań współczesnego *data science*. Celem tego artykułu jest sprawdzenie skuteczności różnych metod rozwiązania problemu nierównowagi danych. W tym celu stworzono kilka modeli klasyfikacji binarnej korzystających ze zbioru danych dotyczących oszustw związanych z kartami kredytowymi. Wykorzystano w nich różne sposoby rozwiązania problemu nierównowagi danych w celu porównania skuteczności klasyfikacji każdego z nich. Została do tego użyta autorska miara skuteczności wykorzystująca wskaźnik F1 modelu oraz czas wykonania kodu.

Słowa kluczowe: resampling, nierównowaga danych, oszustwa związane z kartami kredytowymi

1. Wstęp

W obecnym świecie coraz powszechniej wykorzystuje się płatności przy użyciu kart kredytowych. Terminale płatnicze stają się kluczowym elementem każdego biznesu, w którym dochodzi do regularnych transakcji płatniczych. Tendencja ta spowodowała pojawienie się w XXI wieku nowego rodzaju przestępstw, jakimi są oszustwa z wykorzystaniem karty kredytowej (ang. *credit card fraud*). Charakter tych przestępstw powoduje, że często bywają one trudne do wykrycia. Jednak rozwój technologiczny spowodował, że pomocnym narzędziem przy wykrywaniu tego typu oszustw staje się uczenie maszynowe. Algorytmy *machine learning* są w stanie skutecznie rozpoznać wzorce w danych opisujących użycie konkretnej karty i wykorzystać nabytą w ten sposób wiedzę do skutecznego wykrycia podejrzanych transakcji płatniczych. Jednak istnieje pewien problem, który napotykają na swojej drodze tego typu modele uczenia maszynowego. Dotyczy on zbioru danych, który jest wykorzystywany do ich uczenia i ewaluacji. Przy zadaniu klasyfikacji zbiór powinien być opisany odpowiednimi etykietami sygnalizującymi, czy dana obserwacja dotyczy normalnej płatności, czy też oszustwa. W tym wypadku kłopot wynika z częstotliwości występowania nieprawidłowych transakcji, które stanowią niewielką część wszystkich

operacji dokonywanych za pomocą kart kredytowych. Powoduje to, że zbiory danych dotyczących tego typu transakcji charakteryzują się dużą nierównowagą, która przybiera formę znacznej dominacji przypadków prawidłowych płatności.

Celem tego artykułu jest przedstawienie wyników przeprowadzonego badania związanego ze sprawdzeniem skuteczności zastosowania metod resamplingu danych w przypadku nierównowagi danych przy wykrywaniu oszustw związanych z kartami kredytowymi. Badanie składało się z kilku etapów. Pierwszym było wygenerowanie kilku różnych zestawów danych poprzez zastosowanie metod resamplingu na zbiorze używanym przy wykrywaniu oszustw związanych z kartami kredytowymi. Stworzone w ten sposób zestawy następnie wykorzystano w procesie uczenia kilku modeli klasyfikacji o identycznych parametrach w celu porównania skuteczności ich działania. Do jej określenia użyta została autorska miara będąca średnią arytmetyczną ze znormalizowanych wartości wskaźnika F1 modeli oraz czasu wykonania kodu. Nabyta w ten sposób wiedza powinna pozwolić odpowiedzieć na poniższe pytania badawcze:

1. Jakie są sposoby rozwiązania problemu nierównowagi danych przy wykrywaniu oszustw związanych z kartami kredytowymi?
2. Który z opisanych sposobów jest najskuteczniejszy (co w badanym przypadku oznacza, że charakteryzuje się najwyższą wartością miary skuteczności)?

Struktura artykułu jest następująca: poniżej omówiono rodzaje oszustw związanych z kartami kredytowymi; następnie opisano problem nierównowagi danych; w kolejnym punkcie omówione zostały metody resamplingu danych, a w ostatnim zaprezentowano przebieg, wyniki badania oraz wnioski z niego płynące.

2. Rodzaje oszustw związanych z kartami kredytowymi

Oszustwo związane z kartami kredytowymi można najprościej zdefiniować jako sytuację, w której nieautoryzowana osoba, bez wiedzy posiadacza, uzyskuje dostęp do danych jego karty kredytowej, co daje jej możliwość dokonywania transakcji zakupu za jej pomocą. Z definicji tej jasno wynika, że oszustwem jest również sytuacja, w której nie dochodzi do zmiany posiadacza fizycznej karty. Ten typ określa się mianem *Card-Not-Present Fraud* (CNP) (Abdou i in., 2009). Jest on najpowszechniejszym rodzajem oszustw związanych z kartami kredytowymi, na co wpływa fakt, że jest trudniejszy do wykrycia przez ofiarę, która przez długi czas może pozostawać nieświadoma, że ktoś korzysta z jej karty. Inaczej wygląda sytuacja w przypadku oszustw typu *card-present fraud*, kiedy dochodzi do stworzenia przez oszusta fizycznej kopii karty i posługiwania się nią (Abdou i in., 2009). Ze względu na znaczną wykrywalność rodzaj ten nie cieszy się aktualnie dużą popularnością.

Tematyka oszustw związanych z kartami kredytowymi pojawiła się w literaturze w latach 90. XX wieku. Już wtedy zaczęły się pojawiać pierwsze pomysły na wykorzystanie uczenia maszynowego przy wykrywaniu podejrzanych transakcji płatniczych. Rozwój technologiczny umożliwił nowe sposoby zabezpieczenia kart przed

działaniami oszustów. Jednym z nich było wprowadzenie kodu CVV[1] (ang. *Card Verification Value*). Rolą tego trzyznakowego ciągu cyfr jest zabezpieczenie karty w sytuacjach, gdy do płatności dochodzi bez jej fizycznej obecności (np. w płatnościach internetowych). Zazwyczaj osiąga się to poprzez wymuszenie podania kodu w trakcie realizacji płatności w celu jej uwierzytelnienia.

Ze względu na rodzaj można wyróżnić pięć rodzajów oszustw związanych z kartami kredytowymi:

- Oszustwo upadłościowe (ang. *bankruptcy fraud*) – użycie karty kredytowej w sytuacji, gdy jej posiadacz jest niewypłacalny, przez co nie jest w stanie opłacić zakupionych artykułów czy usług. W takiej sytuacji konieczność zapłaty zobowiązania przechodzi na bank. Najlepszym sposobem na zapobiegnięcie tego typu oszustwom jest przeprowadzenie dokładnej lustracji klienta przed udzieleniem zgody na użycie przez niego karty kredytowej.
- Kradzież (ang. *theft fraud*) – oszust kradnie kartę i wykorzystuje ją tak długo, dopóki nie zostanie ona zablokowana z polecenia jej oryginalnego właściciela. Przy tym rodzaju oszustw kluczowe znaczenie ma szybkość reakcji właściciela karty: im szybciej zorientuje się, że stracił swoją kartę, tym mniejsze szkody zostaną wyrządzone za jej pomocą przez przestępcę.
- Fałszerstwo (ang. *counterfeit fraud*) – zamiast przejęcia fizycznej karty, dochodzi do stworzenia jej kopii przez oszusta, którą następnie wykorzystuje ją przy zakupach, obciążając konto posiadacza oryginalnej karty. Podobnie jak w przypadku kradzieży duże znaczenie ma szybkość reakcji właściciela oryginalnej karty.
- Oszustwo w trakcie aplikowania (ang. *application fraud*) – oszust w momencie starania się o wydanie karty kredytowej składa aplikację zawierającą fałszywe dane. Wyróżnia się dwa rodzaje tego typu oszustw: składanie aplikacji przez posiadacza na te same dane (tzw. tworzenie duplikatu) lub składanie aplikacji na podobne dane (kwalifikowane jako rodzaj fałszerstwa).
- Oszustwo behawioralne (ang. *behavioral fraud*) – oszust pozyskuje dane karty kredytowej, które następnie wykorzystuje przy opłacaniu transakcji niewymagających fizycznej obecności karty. Oszustwo to należy do typu *card-not-present fraud*, ponieważ jego przedmiot stanowią dane karty, a nie ona sama (Abdou i in., 2009).

Wykrywanie oszustw związanych z kartami kredytowymi stanowi poufną dziedzinę ze względu na znaczny udział wrażliwych danych w procesie analizy. Duże znaczenie dla tej dyscypliny mają uczenie maszynowe oraz statystyka, które zapewniają metody umożliwiające skutecznie zdiagnozowanie podejrzanych przypadków, które mogą być przykładami oszustw. Do najpopularniejszych technik wykrywania oszustw należą: sieci neuronowe, drzewa decyzyjne, regresja logistyczna, klasteryzacja oraz wykrywanie wartości odstających (Dubey i in., 2019). Techniki te są powszechnie wykorzystywane przez instytucje finansowe do zabezpieczenia się przed szkodliwymi konsekwencjami oszustw związanych z użyciem kart kredytowych.

[1] Stosowane zamiennie z akronimem CVC (ang. *Card Verification Code*).

3. Problem nierównowagi danych

Problem nierównowagi danych (ang. *imbalance data problem*) można zdefiniować jako sytuację, w której liczebność jednej z klas w zestawie danych jest znacznie większa niż liczebność którejkolwiek z pozostałych klas (Abraham i Elrahman, 2013). Problem ten może dotyczyć zarówno sytuacji, gdy zbiór danych zawiera dwie klasy (zazwyczaj opisane wartościami binarnymi, czyli 0 lub 1), jak i przypadków, gdy liczba klas jest większa. Zależnie od liczebności klas w zestawie danych można przyjąć różne sposoby rozwiązania problemu nierównowagi danych. Kluczowym aspektem jest dobre zrozumienie istoty wykorzystywanego ich zbioru, żeby uniknąć sytuacji, gdy zastosowane rozwiązanie doprowadzi do swoistego zafałszowania informacji, które można z niego wyodrębnić.

W celu zrozumienia źródła nierównowagi danych w pierwszej kolejności należy pojąć, czym jest sam ich zbiór. Zestaw danych (ang. *dataset*) można zdefiniować jako kolekcję dyskretnych, powiązanych ze sobą wartości, które posiadają określoną strukturę oraz dotyczą określonego zagadnienia (Renear i in., 2011). Szczególnie istotny jest ostatni fragment tej definicji, który wskazuje, że dane w pewien sposób opisują jakieś realne zjawisko, starając się przedstawić jego dziedzinę za pomocą atrybutów, które przyjmują określone wartości.

Negatywną konsekwencją tego jest możliwość wystąpienia braku równowagi danych, co najczęściej wynika z dużej różnicy występowania każdego z opisywanych przypadków. Przykładem może być zestaw danych pacjentów wykorzystywany przy klasyfikacji mającej na celu stwierdzenie występowania rzadkiej choroby, np. nowotworu płuc. Mimo że jest to jeden z najpopularniejszych rodzajów nowotworów, niebezpieczeństwo zapadnięcia na tę chorobę ocenia się na poziomie 5-6% (American Cancer Society, 2023). Powoduje to, że znaczna liczba analizowanych obserwacji należy do osób, u których nie stwierdza się obecności tego nowotworu (w przypadku zestawu danych opisanego za pomocą wartości binarnych tego typu obserwacjom zostałaby przypisana wartość 0). Dysproporcja w liczebności obserwacji między klasami dla tego przypadku może osiągać wartości na poziomie nawet 25:1 (Alam i in., 2021).

W przypadku zbioru danych zawierającego wiele klas interesującą metodą na rozwiązanie problemu nierównowagi danych jest zdefiniowanie nowych klas poprzez „połączenie" istniejących. Sposób ten jest stosunkowo prosty i może się okazać skuteczny, jednak posiada dwie zasadnicze wady, które sprawiają, że jego zastosowanie nie zawsze jest możliwe:

- niewłaściwe połączenia klas mogą zafałszować zbiór danych,
- skrajne klasy mogą być na tyle mało liczne, że nawet po połączeniu mogą zawierać zdecydowanie mniejszą liczbę obserwacji niż pozostałe (Krawczyk, 2016).

Oba opisane problemy można skutecznie zilustrować na przykładzie zbioru danych wykorzystywanego przy analizie ryzyka kredytowego. Najważniejszym aspektem, który bierze się pod uwagę przy podejmowaniu decyzji dotyczącej udzielenia

kredytu, jest rating kredytowy danego podmiotu. Jest to ogólna ocena zdolności kredytowej wyrażana za pomocą oznaczeń literowych, zazwyczaj w skali od AAA do D, gdzie AAA jest najwyższą możliwą oceną (White, 2010). Skale ocen największych agencji ratingowych zostały zaprezentowane na rysunku 1.

S&P	Moody's	Fitch
AAA	Aaa	AAA
AA+	Aa1	AA+
AA	Aa2	AA
AA-	Aa3	AA-
A+	A1	A+
A	A2	A
A-	A3	A-
BBB+	Baa1	BBB+
BBB	Baa2	BBB
BBB-	Baa3	BBB-
BB+	Ba1	BB+
BB	Ba2	BB
BB-	Ba3	BB-
B+	B1	B+
B	B2	B
B-	B3	B-
CCC+	Caa1	CCC
CCC	Caa2	
CCC-	Caa3	
CC	Ca	CC
C	C	C
D		D

Rys. 1. Skale ocen największych agencji ratingowych

Źródło: opracowanie własne na podstawie (Basar i Genc, 2019).

W przypadku ratingów agencji S&P i Fitch jednym ze sposobów na „połączenie" klas mogłoby być użycie pierwszej litery oceny. W ten sposób powstałyby cztery klasy (A, B, C i D). Jednak należy wziąć pod uwagę, że zazwyczaj niewielki procent podmiotów otrzymuje najniższe oceny, w związku z czym w takim układzie liczba obserwacji należących do klas C i D byłaby znacznie niższa niż w przypadku dwóch pozostałych klas, przez co zestaw danych w dalszym ciągu byłby niezbalansowany.

Innym problemem tego typu łączenia jest to, że zazwyczaj granica decyzyjna dotycząca udzielenia kredytu danemu podmiotowi przebiega między klasami BBB- oraz BB+, przez co połączenie obserwacji należących do klas zaczynających się na literę B sprawiłoby, że zbiór zostałby zafałszowany, ponieważ jedynie dla części obserwacji z tej nowo powstałej klasy decyzja byłaby pozytywna. Rozwiązaniem tego problemu mogłoby być wydzielenie większej liczby klas (np. AAA, AA, A, BBB, BB, B, CCC, C oraz D). Taki podział pozwoliłby uniknąć zafałszowania danych, jednak duża liczba klas mogłaby spowodować, że w dalszym ciągu występowałaby dysproporcja w liczbie obserwacji między najbardziej skrajnymi spośród nich (w tym przypadku AAA oraz CCC, C i D) oraz pozostałymi.

W przypadku zestawów danych wykorzystywanych w obu rodzajach klasyfikacji (binarnej i wieloklasowej) za najskuteczniejsze metody rozwiązania problemu nierównowagi danych uznaje się:

- metody poziomu danych (ang. *data-level methods*) – techniki rozpoziomowania (resamplingu) danych,
- metody poziomu algorytmów (ang. *algorithm-level methods*) – techniki, w ramach których standardowe algorytmy uczenia maszynowego są przystosowywane do skutecznego przetwarzania niezbalansowanych zbiorów danych poprzez np. zastosowanie zmiennych określających wagę lub koszt,
- metody hybrydowe – kombinacja metod poziomu danych i poziomu algorytmów (Alam i in., 2021).

Spośród wymienionych technik największą popularnością cieszy się resampling danych. W znacznym stopniu wpływa na to różnorodność wariantów, których można użyć w celu zmiany liczebności obserwacji niezbalansowanych klas.

4. Metody resamplingu danych

Resampling danych jest metodą poziomu danych wykorzystywaną do zmiany liczby obserwacji w niezbalansowanych klasach. Techniki resamplingu stosuje się na etapie wstępnego przetwarzania (ang. *pre-processing stage*). Ich użycie sprawia, że struktura zbioru danych zostaje zmieniona w celu skutecznego poradzenia sobie z problemem nierównowagi danych. Tego typu działanie do pewnego stopnia przyczynia się do poprawy zdolności predykcyjnych modelu nauczonego na podstawie resamplowanego zbioru danych (Lee, 2014).

W ramach resamplingu wyróżnia się trzy[2] różne podejścia, które można zastosować do zmiany liczebności obserwacji. Są to:

- *undersampling* – rodzaj resamplingu, w ramach którego liczba obserwacji należących do bardziej licznych klas jest redukowana do poziomu najmniej licznej klasy przy zachowaniu istotnych informacji z perspektywy procesu uczenia modelu,
- *oversampling* – rodzaj resamplingu, w ramach którego liczba obserwacji należących do mniej licznych klas jest zwiększana do poziomu najbardziej licznej klasy przy zachowaniu istotnych informacji z perspektywy procesu uczenia modelu,
- *sampling* hybrydowy – rodzaj resamplingu, w ramach którego dochodzi do połączenia metod undersamplingu i oversamplingu, co ma na celu ograniczenie negatywnych konsekwencji zastosowania tylko jednej z nich (Alam i in., 2021).

Oczywiste jest, że żaden z wymienionych rodzajów nie jest idealny. Największy problem undersamplingu stanowi fakt, że zastosowanie jego metod sprawia, iż część obserwacji zawierających istotne informacje może zostać usunięta ze zbioru danych, co negatywnie wpływa na jego jakość. Różnorodność zasad filtrowania w ramach różnych metod undersamplingu ma w założeniu ograniczać występowanie tego typu negatywnego odrzucenia ważnych obserwacji, jednak w takim wypadku pojawia się kolejny problem, jakim jest wybranie odpowiedniej metody dla posiadanych danych.

[2] Wobec oversamplingu i undersamplingu używa się również określeń upsampling i downsampling.

W przypadku oversamplingu nowe obserwacje są tworzone na podstawie istniejących w celu osiągnięcia bardziej zbalansowanej ich dystrybucji między klasami. Negatywnym skutkiem takiego działania jest możliwość występowania zjawiska przeuczenia (ang. *overfitting*) modelu. W uczeniu maszynowym przeuczenie (zwane również nadmiernym dopasowaniem) jest sytuacją, w której model zbyt dobrze „dopasowuje się" do danych treningowych. Powoduje to, że jego zdolność do prognozowania na podstawie niewidzianych wcześniej danych ze zbioru testowego jest osłabiona, przez co *de facto* zatraca on istotę swojego funkcjonowania, którą jest tworzenie przewidywań na podstawie nieznanych danych. Ryzyko przeuczenia wzrasta, gdy zbiór treningowy jest zbyt mały pod względem liczby obserwacji. Tak jest w przypadku zastosowania metod oversamplingu, ponieważ po ich wykorzystaniu liczba obserwacji dla mniej licznej klasy może się wydawać duża, jednak w znacznej mierze są to sztuczne obserwacje, wygenerowane na podstawie niewielkiego zbioru istniejących. Powoduje to, że proces uczenia modelu odbywa się faktycznie na oryginalnej próbce zwiększonej przez pochodne należących do niej obserwacji. Efektem tego jest możliwość wystąpienia zjawiska przeuczenia.

Dla każdego z podejść można wyróżnić wiele metod resamplingu, które wykorzystują różne podejścia do zmiany liczebności obserwacji w zbiorze danych. Przykładowe metody zostały zaprezentowane w tabeli 1.

Tabela 1. Przykładowe metody resamplingu

Undersampling	*Oversampling*
Random Under-Sampling (RUS)	*Random Over-Sampling* (ROS)
Edited Nearest Neighbors (ENN)	*Adaptive Synthetic* (ADASYN)
Tomek Links (TL)*	*Synthetic Minority Over-Sampling Technique* (SMOTE)

* Nazwa metody pochodzi od nazwiska jej twórcy – Ivana Tomka.

Źródło: opracowanie własne na podstawie (Alam i in., 2021).

Spośród wymienionych w tabeli 1 przykładów metody RUS i ROS działają na podobnej zasadzie poprzez wykorzystanie losowości do resamplingu danych. RUS usuwa losowo wybrane obserwacje z bardziej licznych klas. Tymczasem ROS zwiększa liczebność mniej licznych klas poprzez kopiowanie istniejących obserwacji. Obie metody są najmniej skomplikowanymi metodami resamplingu, jednak ich prostota wiąże się z tym, że charakteryzują się mniejszą skutecznością niż inne.

Edited Nearest Neighbors

Podstawę tej metody stanowi algorytm *k* najbliższych sąsiadów (ang. *K-Nearest Neighbors* – KNN), który przyporządkowuje obserwacje do klas na podstawie przy-

należności klasowej ich k[3] najbliższych sąsiadów. W ramach metody ENN każda z obserwacji należących do bardziej licznych klas jest testowana za pomocą KNN. Obserwacje, które zostały błędnie sklasyfikowane przez KNN, zostają następnie usunięte ze zbioru danych (Alam i in., 2021).

Tomek Links

Podstawą działania tej metody jest zidentyfikowanie wśród obserwacji bardziej licznej klasy, tzw. *Tomek Links*, czyli par przypadków, które spełniają dwa warunki:

- należą do różnych klas,
- są swoimi najbliższymi sąsiadami.

W przypadku skutecznego zidentyfikowania, obserwacja, która jednocześnie jest jednym z *Tomek Links* oraz należy do bardziej licznej klasy, jest usuwana ze zbioru danych. W ten sposób dochodzi do undersamplingu poprzez usunięcie przypadków brzegowych (ang. *boundary instances*) (Alam i in., 2021).

Adaptive Synthetic

Adaptive Synthetic to metoda oversamplingu wykorzystująca rozkład ważony (ang. *weighted distribution*) mniej licznej klasy do wygenerowania syntetycznych danych, czyli sztucznych obserwacji wykorzystywanych do poprawy działania modelu uczenia maszynowego. W ramach tej metody dla każdej obserwacji należącej do mniej licznej klasy wykorzystuje się algorytm KNN w celu zdefiniowania jej najbliższego sąsiedztwa (ang. *neighborhood*). Następnie dochodzi do wyliczenia znormalizowanego współczynnika dominacji bardziej obfitej klasy dla każdego z sąsiedztw. Jest on wykorzystywany do określenia liczby przypadków mniej licznej klasy, które należy wygenerować dla każdego z sąsiedztw w celu osiągnięciu identycznej liczebności wszystkich klas.

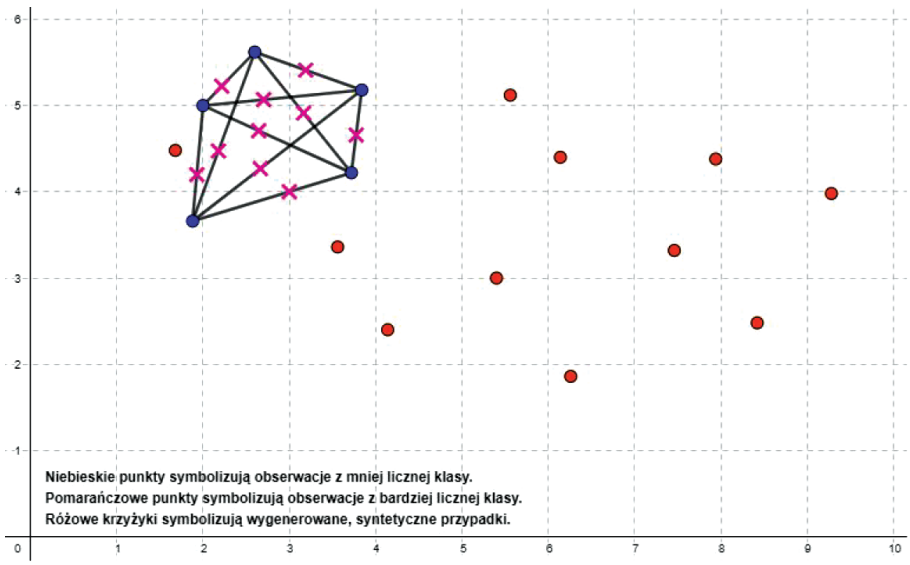
Nowe obserwacje dla skąpszej klasy są generowane na podstawie liniowej kombinacji dwóch losowo wybranych obserwacji, które należą do tej klasy oraz znajdują się w obrębie tego samego sąsiedztwa. Częstym zabiegiem jest dodawanie do wygenerowanych przypadków tzw. białego szumu (ang. *white noise*) w celu uniknięcia zbyt dużej liniowej zależności między obserwacjami w ramach klasy, która jest poddawana oversamplingowi (Alam i in., 2021).

Synthetic Minority Over-Sampling Technique

W metodzie SMOTE syntetyczne dane dla mniej licznej klasy są generowane poprzez interpolację k najbliższych sąsiadów dla każdej z obserwacji należącej do tej klasy. Schemat działania metody SMOTE jest bardzo podatny, jak w przypadku *Adaptive Synthetic*, z tą istotną różnicą, że lokalny rozkład klasy mniejszościowej nie wpływa na liczbę przypadków generowanych dla każdego z sąsiedztw. Powoduje to, że implementacja SMOTE jest prostsza.

[3] Litera „k" symbolizuje liczbę najbliższych sąsiadów, którą chcemy wziąć pod uwagę, w związku z czym można ją zastąpić dowolną liczbą naturalną.

Pierwszym krokiem w ramach tej metody jest losowe wybranie jednej z obserwacji zaliczanej do mniej licznej klasy, dla której należy znaleźć *k* najbliższych sąsiadów. Następnie należy wybrać jednego z tych sąsiadów, którzy jednocześnie również należą do skąpszej klasy. Kolejnym krokiem jest zdefiniowanie syntetycznego punktu na linii, która łączy obie wybrane obserwacje. Każdy taki punkt jest wygenerowanym przypadkiem klasy, która jest poddawana oversamplingowi. W związku z tym trzeba powtarzać ten proces do momentu, gdy liczebność klas będzie zbalansowana. Przykład zastosowania metody SMOTE został zaprezentowany na rysunku 2.



Rys. 2. Przykład zastosowania metody SMOTE

Źródło: opracowanie własne.

Hybrydowe podejście do resamplingu łączy metody oversamplingu i undersamplingu w celu zredukowania negatywnego wpływu wad obu podejść. Częstym połączeniem metod w ramach resmaplingu hybrydowego jest jednoczesne użycie metod SMOTE i ENN, a także SMOTE i *Tomek Links* (Alam i in., 2021).

5. Porównanie metod resamplingu danych

5.1. Założenia realizacji badania

Rozwój technologiczny sprawia, że liczba źródeł generowania danych systematycznie rośnie. Wykorzystanie technologii Internetu rzeczy umożliwia szybkie przesyłanie danych między urządzeniami w celu ich przechowywania oraz przetwarzania.

Przykładem takiego generatora danych są terminale płatnicze, które przy obsłudze transakcji zbierają informacje na jej temat, wykorzystywane później do wykrywania oszustw przy płatnościach bezgotówkowych i zapobiegania im.

Na potrzeby przeprowadzenia tego studium przypadku wybrany został zbiór danych *Credit Card Fraud Detection*, dostępny na stronie internetowej Kaggle. Zbiór ten zawiera dane dotyczące transakcji przy użyciu kart kredytowych, które zostały wykonane na przestrzeni dwóch dni we wrześniu 2013 roku. Ze względu na kwestię poufności danych większość spośród 31 atrybutów stanowi 28 cech będących głównymi składowymi wyodrębnionymi w ramach ich analizy dokonanej na oryginalnych danych. Pozostałe trzy atrybuty to: czas dokonania transakcji, jej kwota oraz klasa w formacie binarnym przyjmująca wartość „0" dla prawdziwych transakcji oraz „1" dla fałszywych. Zbiór składa się z 284 807 wierszy, przy czym każdy z nich odpowiada pojedynczej transakcji płatniczej. Zestaw danych jest mocno niezbalansowany, o czym najlepiej świadczy fakt, że do klasy „1" należą tylko 492 wiersze, co stanowi zaledwie 0,17% całości.

Badanie przeprowadzono według następującej procedury:

1. Wybór danych dotyczących użycia kart kredytowych.
2. Wygenerowanie siedmiu różnych zestawów danych poprzez zastosowanie różnych metod resamplingu na wybranym zbiorze danych.
3. Stworzenie siedmiu różnych modeli klasyfikatora lasu losowego o identycznych parametrach.
4. Wytrenowanie modeli na podstawie przyporządkowanych zestawów danych.
5. Ewaluacja modeli w celu uzyskania wartości wykorzystywanych do obliczenia miary skuteczności dla każdego z modeli (wskaźnika F1 oraz czasu wykonania kodu).
6. Normalizacja uzyskanych wartości.
7. Obliczenie wartości miary skuteczności dla każdej z metod resamplingu danych na podstawie znormalizowanych wartości wskaźnika F1 i czasu wykonania kodu.

5.2. Realizacja badania oraz otrzymane wyniki

W celu sprawdzenia, która z metod resamplingu danych jest najskuteczniejsza, przeprowadzone zostały poniżej opisane czynności. W pierwszej kolejności doszło do wytrenowania modelu klasyfikatora lasu losowego (ang. *Random Forest Classifier* – RFC) przy użyciu zbioru uczącego z oryginalnego, niezbalansowanego zestawu danych. Po zakończeniu procesu uczenia modelu został on poddany ewaluacji poprzez obliczenie dla niego czasu wykonania wszystkich operacji (ewentualnego resamplingu, tworzenia instancji modelu, jego uczenia i dokonania przez niego predykcji oraz obliczenia metryk ewaluacyjnych) w środowisku *Google Colab* oraz wskaźnika F1, który oblicza się za pomocą precyzji (ang. *precision*) i wrażliwości (ang. *recall*).

$$F1 = 2 \times \frac{Precyzja \times Wrażliwość}{Precyzja + Wrażliwość}.$$

Wskaźnik F1 zaleca się wykorzystywać przy ewaluacji modelu klasyfikacji wykorzystującego niezbalansowany zbiór danych, ponieważ powszechniej stosowana dokładność (ang. *accuracy*) w tego typu przypadkach może dawać zafałszowaną wartość. Wynika to z faktu, że jest to miara określająca: jaka część wszystkich obserwacji została poprawnie zakwalifikowana, co oznacza, że przy niezbalansowanych zbiorach danych jej wartość może być sztucznie zawyżana przez poprawnie zaklasyfikowane wartości należące do bardziej licznej klasy (Bej i in., 2021).

Ogólne wyniki ewaluacji modelu wyuczonego na oryginalnym zestawie danych zostały zaprezentowane w tabeli 2.

Tabela 2. Wyniki ewaluacji modelu wyuczonego na oryginalnym zestawie danych

Nazwa metryki	Dokładność	Precyzja	Wrażliwość	Wskaźnik F1
Wartość	0,999607	0,898990	0,787611	0,839622

Źródło: opracowanie własne.

W następnych krokach doszło do powtórzenia wyżej opisanych czynności dla każdej spośród sześciu analizowanych metod resamplingu danych (SMOTE, ADASYN, ENN, *Tomek Links*, SMOTE-ENN, SMOTE-*Tomek Links*). Po dokonaniu zmiany liczby obserwacji w zbiorze w ramach każdej z metod, zostały one wykorzystane w procesie uczenia modeli RFC. Dla każdej z metod resamplingu powstał oddzielny model, który został poddany ewaluacji poprzez obliczenie czasu wykonania wszystkich operacji w środowisku *Google Colab* oraz wskaźnika F1.

Wartości czasu wykonania oraz wskaźnika F1 dla każdego z modeli wykorzystującego dane, dla których zostały zastosowane badane metody resamplingu, zaprezentowano w tabeli 3.

Tabela 3. Wyniki ewaluacji modeli wyuczonych na danych, dla których zastosowane zostały badane metody resamplingu

Nazwa metody resamplingu	Wartość wskaźnika F1	Czas wykonania [s]
Brak resamplingu	0,839622	281,05
Edited Nearest Neighbors	0,865672	806,48
Tomek Links	0,878924	814,11
Adaptive Synthetic	0,999894	560,90
SMOTE	0,999902	572,13
SMOTE-ENN	0,999958	2601,51
SMOTE-*Tomek Links*	0,999880	2646,35

Źródło: opracowanie własne.

Analiza wyników pozwala na sformułowanie wniosku, że w danym przypadku najwyższe wartości wskaźnika F1 osiągają metody oversamplingu oraz samplingu hybrydowego. Pod względem czasu wykonania należy wyróżnić metody oversamplingu,

które charakteryzowały się znacznie krótszym czasem niż inne analizowane metody. W celu wykrycia najskuteczniejszego sposobu resamplingu dla danego przypadku wyniki dla obu wybranych kryteriów oceny zostały poddane normalizacji min-max. Wyniki ewaluacji modeli po normalizacji zostały zaprezentowane w tabeli 4.

Tabela 4. Wyniki ewaluacji modeli po normalizacji

Nazwa metody resamplingu	Wartość wskaźnika F1	Czas wykonania
Brak resamplingu	0	1
Edited Nearest Neighbors	0,162471	0,777859
Tomek Links	0,245123	0,774633
Adaptive Synthetic	0,999601	0,881685
SMOTE	0,999651	0,876937
SMOTE-ENN	1	0,018957
SMOTE-*Tomek Links*	0,999514	0

Źródło: opracowanie własne.

Po normalizacji wyników dla każdej z metod wyliczona została miara skuteczności stanowiąca średnią arytmetyczną ze znormalizowanych wartości wskaźnika F1 oraz czasu wykonania:

$$\text{miara skuteczności} = \frac{F1'+\text{czas}'}{2} \cdot$$

Tabela 5. Wartości miary skuteczności dla modeli wyuczonych na danych, dla których zastosowane zostały badane metody resamplingu

Nazwa metody resamplingu	Wartość miary skuteczności
Brak resamplingu	0,500000
Edited Nearest Neighbors	0,470166
Tomek Links	0,509878
Adaptive Synthetic	0,940643
SMOTE	0,938294
SMOTE-ENN	0,509479
SMOTE-*Tomek Links*	0,499756

Źródło: opracowanie własne.

Najskuteczniejsza metoda będzie charakteryzowała się najwyższą wartością tej miary. Wartości miary skuteczności zostały zaprezentowane w tabeli 5.

5.3. Wnioski z przeprowadzonego badania

Spośród wszystkich analizowanych metod najlepszy wynik miary skuteczności uzyskano w przypadku *Adaptive Synthetic* (0,940643). Na drugim miejscu znalazł się *Synthetic Minority Oversampling Technique* z wartością miary skuteczności na poziomie 0,938294. Dwie metody (*Edited Nearest Neighbors* oraz SMOTE-Tomek Links) uzyskały wartości na niższym poziomie niż w przypadku braku resamplingu. W związku z tym brak jest przesłanek do wykorzystania ich na badanym zbiorze danych.

6. Zakończenie

W niniejszym artykule przedstawiono wyniki oraz wnioski z przeprowadzonego badania mającego sprawdzić, która spośród wybranych metod resamplingu danych jest najskuteczniejsza przy zbiorach danych dotyczących wykrywania oszustw związanych z kartami kredytowymi. W tym przypadku skuteczność została wyrażona jako średnia arytmetyczna między znormalizowanymi wartościami wskaźników F1 modeli oraz czasów wykonania całości kodu w środowisku *Google Colab*, na który składały się: resampling oraz podział danych, stworzenie instancji modelu klasyfikatora lasu losowego, przeprowadzenie procesu jego uczenia, dokonanie przez niego predykcji na podstawie zbioru testowego oraz wyliczenie metryk ewaluacyjnych dla predykcji.

Wybrane kryterium oceny jednoznacznie wskazało najskuteczniejsze metody oversamplingu. Najlepsze wyniki miary skuteczności zostały uzyskane przez *Adaptive Synthetic* oraz *Synthetic Minority Oversampling Technique*. Pozostałe badane metody osiągnęły wyniki bliskie wynikowi uzyskanemu przez oryginalny zestaw danych.

W ramach przyszłych publikacji można by rozszerzyć zakres przeprowadzonych badań poprzez sprawdzenie większej liczby metod resamplingu dla każdego z rodzajów. Dodatkowo warto byłoby również przeprowadzić badanie wśród pracowników przedsiębiorstw z rynku kart płatniczych w celu ustalenia, który z aspektów zaproponowanej miary skuteczności byłby dla nich istotniejszy (wartość wskaźnika F1 lub czas wykonania kodu). Uzyskane wyniki mogłyby pomóc w nadaniu odpowiednich wag obu składowym miary, które można by wykorzystać do przeprowadzenia kolejny raz tego samego badania, żeby porównać wyniki dla oryginalnej miary z tymi dla nowej wersji uwzględniającej istotność z perspektywy użytkowników.

Literatura

Abdou, H., Delamaire, L. i Pointon, J. (2009). Credit Card Fraud and Detection Techniques: a Review. *Banks and Bank Systems, 4*(2), 57-68.

Abraham, A. i Elrahman, S. (2013). A Review of Class Imbalance Problem. *Journal of Network and Innovative Computing, 1*(2013), 332-340.

Alam, T. M., Hameed, I. A., Khushi, M., Luo, S., Reyes, M. C., Shaukat, K., … S., Yang, X. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE,* (9), 109960-109975.

American Cancer Society. (2023). *Key Statistics for Lung Cancer*. American Cancer Society.

Basar, O. D. i Genc, E. G. (2019). Comparison of Country Ratings of Credit Rating Agencies with MOORA Method. *Business and Economics Research Journal, 10*(2), 391-404.

Bej, S., Davtyan, N., Nassar, M., Wolfien, M. i Wolkenhauser, O. (2021). LoRAS: An Oversampling Approach for Imbalanced Datasets. *Machine Learning, 110*, 279-301.

Dubey, S., Jain, S., Jain, Y. i Tiwari, N. (2019). A Comparative Analysis of Various Credit Card Fraud Detection Techniques. *IJRTE*, *7*(5S2), 402-407.

Krawczyk, B. (2016). Learning from Imbalanced Data: Open Challenges and Future Directions. *Progress in Artificial Intelligence, 5*, 221-232.

Lee, P. H. (2014). Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets. *Int J Environ Res Public Health*, *11*(9), 9776-9789.

Renear, A., Sacchi, S. i Wickett, K. (2011). Definitions of Dataset in the Scientific and Technical Literature. *ASIS&T*, *47*(1), 1-4.

White, L. (2010). Markets: The Credit Rating Agencies. *Journal of Economic Perspectives*, *24*(2), 211-226.

Means of Addressing Data Imbalance in Credit Card Fraud Detection

Abstract: Data imbalance is one of the most common problems in modern days classification tasks. In machine learning, the data sample is supposed to be a reliable representation of the population as a whole. However, it is equally important that it is constructed in such a way as to provide the model with the best conditions during the learning process. Finding a balance between these two aspects is one of the challenges of modern data science. The purpose of this article is to test the effectiveness of various techniques for solving the data imbalance problem. To achieve this, several binary classification models were created using a dataset dealing with credit card fraud. The models used different ways of solving the data imbalance problem in order to compare the effectiveness of each model's classification. A proprietary performance measure was used to determine the effectiveness of the models, using the models' F1 scores and code execution times.

Keywords: resampling, data imbalance, credit card fraud