

Wpływ wrogiego uczenia maszynowego na modele estymacji wiarygodności kredytowej

Mateusz Kuchta

4 sierpnia 2023

Spis treści

Wstęp	5
1 Credit Scoring	6
1.1 Credit Scoring - definicja i cechy	6
1.2 Wyznaczanie oceny punktowej	10
1.3 Score kredytowy w erze Big Data	15
1.4 Metody Uczenia Maszynowego wykorzystywane w dziedzinie Credit Scoring	21
2 Adversarial Machine Learning – Wrogie Uczenie Maszynowe	26
2.1 Wprowadzenie do technik Uczenia Maszynowego	26
2.2 Charakterystyka Wrogiego Uczenia Maszynowego	30
2.3 Typy ataków na algorytmy Uczenia Maszynowego	34
2.4 Przykłady ataków na algorytmy Uczenia Maszynowego	39
3 Budowa modelu drzewa decyzyjnego do celu Credit Scoring	46
3.1 Opis zbioru danych oraz użytych narzędzi i technologii	46
3.1.1 Tytuł podrozdziału	46
3.1.2 Tytuł podrozdziału	46
3.1.3 Tytuł podrozdziału	46
3.2 Budowa modelu	47
3.2.1 Tytuł podrozdziału	47
3.2.2 Tytuł podrozdziału	47
3.2.3 Tytuł podrozdziału	47
3.3 Analiza wyników	48
3.3.1 Tytuł podrozdziału	48
3.3.2 Tytuł podrozdziału	48
3.3.3 Tytuł podrozdziału	48
4 Atak na opracowany model	49
4.1 Strategia badania odporności modelu	49
4.1.1 Tytuł podrozdziału	49
4.1.2 Tytuł podrozdziału	49
4.1.3 Tytuł podrozdziału	49

4.2	Implementacja wybranych technik ataku	50
4.2.1	Tytuł podpodrozdziału	50
4.2.2	Tytuł podpodrozdziału	50
4.2.3	Tytuł podpodrozdziału	50
4.3	Analiza wyników i weryfikacja hipotez badawczych	51
4.3.1	Tytuł podpodrozdziału	51
4.3.2	Tytuł podpodrozdziału	51
4.3.3	Tytuł podpodrozdziału	51
	Wnioski	52
	Literatura	53
	Spis rysunków	59
	Spis tabel	60
	Załączniki	61

Wstep

1 Credit Scoring

Ocena wiarygodności kredytowej jest stosowana na całym świecie do przetwarzania wielu rodzajów pożyczek. Jest ona wykorzystywana najszerzej i z największym powodzeniem w przypadku osobistych kart kredytowych oraz kredytów hipotecznych. Ryzyko spłaty tych zobowiązań jest ściśle powiązane z czynnikami weryfikowalnymi, takimi jak dochód, ocena Biura Informacji Kredytowej, czy też demografia, tj. wiek, wykształcenie, status cywilny itp. (Caire, Barton, de Zubiria, Alexiev, & Dyer, 2006). Nieraz trudno jest ocenić, czy dana osoba zasługuje na zaufanie, czy może tym razem bank powinien się wstrzymać, nie ryzykując problemami ze spłatą danego kredytobiorcy, jednocześnie rezygnując z potencjalnego zysku.

Dokładne określenie granicy między dobrym, a złym klientem jest trudne nawet dla najbardziej doświadczonych pracowników finansowych. Zwiększona konkurencja i rosnąca presja na generowanie przychodów skłoniły instytucje udzielające kredytów do poszukiwania skutecznych sposobów pozyskiwania nowych klientów, przy jednoczesnej kontroli zysków i strat. Agresywne działania marketingowe wymusiły konieczność dokładniejszej analizy danych potencjalnych klientów, a potrzeba szybkiego i efektywnego ich przetwarzania doprowadziła do rosnącej automatyzacji procesu składania wniosków kredytowych i ubezpieczeniowych, a co za tym idzie, skrócenia czasu ich rozpatrywania (Siddiqi, 2016).

1.1 Credit Scoring - definicja i cechy

Credit Scoring można zdefiniować jako zespół metod statystycznych, używany w celu wyznaczania prawdopodobieństwa nie wywiązania się wnioskodawcy ze spłaty zaciągniętych zobowiązań w ustalonym terminie, co pomaga ustalić, czy kredyt powinien być przyznany potencjalnemu kredytobiorcy. Scoring kredytowy jest jedną z metod systematycznej oceny, która została uznana za istotnie wpływającą na obniżenie poziomu ryzyka wygenerowania strat finansowych w bankach.

W literaturze można znaleźć wiele definicji scoringu, co wynika z faktu, że jest to pojęcie w znacznym stopniu subiektywne. Ważne jest, aby model scoringowy charakteryzował się wysoką skutecznością w oddzielaniu klientów przynoszących zyski od tych, którzy prawdopodobnie przyniosą straty (Wysiński, 2013). Jest to zatem kluczowe narzędzie w rękach instytucji finansowych, dające możliwość ograniczania potencjalnego ryzyka współ-

pracy z niewiarygodnym klientem, przy jednoczesnym osiąganiu jak największych korzyści finansowych poprzez zawieranie umów z wartościowymi kredytobiorcami(direct.money.pl, 2022).

Credit Scoring charakteryzuje się kilkoma podstawowymi cechami(mfiles.pl, 2020):

- Dane historyczne jako baza – porównuje się ze sobą charakterystyki grup kredytobiorców rzetelnych z nierzetelnymi i na tej podstawie dokonuje się oceny, czy potencjalny klient będzie terminowo spłacał zaciągnięte zobowiązanie, czy też może istnieje wysokie prawdopodobieństwo, że zachowa się on podobnie jak usługobiorcy mający problemy z uiszczaniem kolejnych rat na czas;
- Okresowość – mechanizmy wyliczania oceny wymagają częstych aktualizacji o nowe dane, w celu zapewnienia jak najwyższej skuteczności otrzymanego wyniku;
- Rzetelne zbadanie zdolności kredytowej kredytobiorców jako środek do celu jakim jest ochrona interesów kredytodawcy;
- Realizowany za pomocą zaakceptowanych i zatwierdzonych metod statystycznych.

Zwykle mówi się o dwóch typach scoringu - aplikacyjnym i behawioralnym. Pierwszy z nich jest stosowany u klientów, o których informacje są dostępne jedynie na podstawie wypełnionych przez nich wniosków kredytowych oraz danych pozyskanych z zewnętrznych źródeł np. z BIK - Biura Informacji Kredytowej.

Scoring behawioralny bierze pod uwagę informacje zgromadzone podczas wcześniejszej współpracy z klientem. Stanowi również narzędzie wspomagające monitorowanie portfela kredytowego oraz ograniczanie wysokości rezerw tworzonych na tzw. kredyty zagrożone. System scoringu behawioralnego jest szczególnie przydatny przy określaniu nowego limitu kredytowego lub modyfikacji limitu przyznanego wcześniej. Stosuje się go również w celu przeciwdziałania przekraczaniu określonych przez bank limitów na rachunkach, czy przedłużaniu warunków umów na dodatkowe produkty (np. kartę kredytową). Zatem podstawową różnicę pomiędzy scoringiem aplikacyjnym, a behawioralnym stanowi grupa docelowa klientów(Matuszyk, 2009).

W analizie ryzyka przedsiębiorstw stosuje się inny typ oceny, tzw. profit scoring (ang. scoring zysku). Taki system punktowania pozwala na określenie maksymalnego zysku, zamiast minimalizacji ryzyka związanego z obsługą danego klienta. Profit scoring, jako

rozszerzenie podstawowego modelu scoringowego, bierze pod uwagę szereg dodatkowych czynników ekonomicznych tj. strategie marketingowe, dobór polityki cenowej, czy też poziom obsługi (bankier.pl, 2007).

Punktowa ocena wiarygodności kredytowej budowana jest na zasadzie przyznawania punktów za określone cechy kredytobiorcy, gdzie im wyższy wynik wnioskodawca osiągnie, tym większa szansa, że spłaci kredyt w terminie. Tabela punktowa tworzona jest na podstawie analizy statystycznej bazy danych klientów z przeszłości, gdzie poszukuje się cech, które w jak najlepszy sposób oddzielają od siebie dobrych i złych biorców kredytowych (bankier.pl, 2012).



Rysunek 1: Przykład cechy wykorzystanej w Credit Scoring (bankier.pl, 2012)

Dla pewnego modelu, wpływ posiadania własnego mieszkania na spłacenie zobowiązania bez opóźnień zwizualizowano na rysunku 1. Widać na nim, że wśród osób mających problemy ze spłatą pożyczki (200 klientów), aż 95 procent (190 klientów) stanowią osoby wynajmujące nieruchomość. Zatem posiadanie własnego mieszkania będzie stawiać w uprzywilejowanej pozycji potencjalnych klientów banku i otrzymają oni wyższą ocenę za tę cechę w ogólnym scoring'u. Taką logiką kierują się zarówno banki, jak i BIK. Sposób wyliczania oceny punktowej często jest tajemnicą w instytucjach finansowych i zwykle nie dowiemy się jaki score otrzymaliśmy oraz co jest powodem takiego wyniku.

W Biurze Informacji Kredytowej, mimo iż nie poznamy pełnej logiki oceniania, pośrednio wiadome są najważniejsze kryteria oszacowań. Wśród najważniejszych z nich należy wymienić (bankier.pl, 2012):

- Liczba nieterminowo spłacanych zobowiązań;
- Wysokość zobowiązań (np. na karcie kredytowej lub debetowej);
- Czas zwłoki ze spłatą zobowiązania;
- Czas jaki upłynął od ostatniego wykroczenia.

Nieodłącznym elementem predykcji zachowań potencjalnych kredytobiorców względem zaciągniętego zobowiązania jest szacowanie zdolności kredytowej. O ile w przypadku kredytów gotówkowych banki często stosują dosyć liberalne podejście, to gdy pod uwagę brane są duże kwoty pożyczki, tak jak ma to miejsce w przypadku kredytów hipotecznych, etap estymacji zdolności stanowi gęste sito dla wnioskujących, szczególnie w czasach kryzysów gospodarczych, gdzie najczęściej mamy do czynienia z wysokimi stopami procentowymi. Oczywiście analiza zdolności kredytowej opiera się na różnych kryteriach, które dodatkowo różnią się między poszczególnymi bankami, jednak relacja zarobków do wydatków, które są pośrednio zależne od poziomu stóp procentowych, stanowi bazę do dalszej oceny (habza.com.pl, 2022).

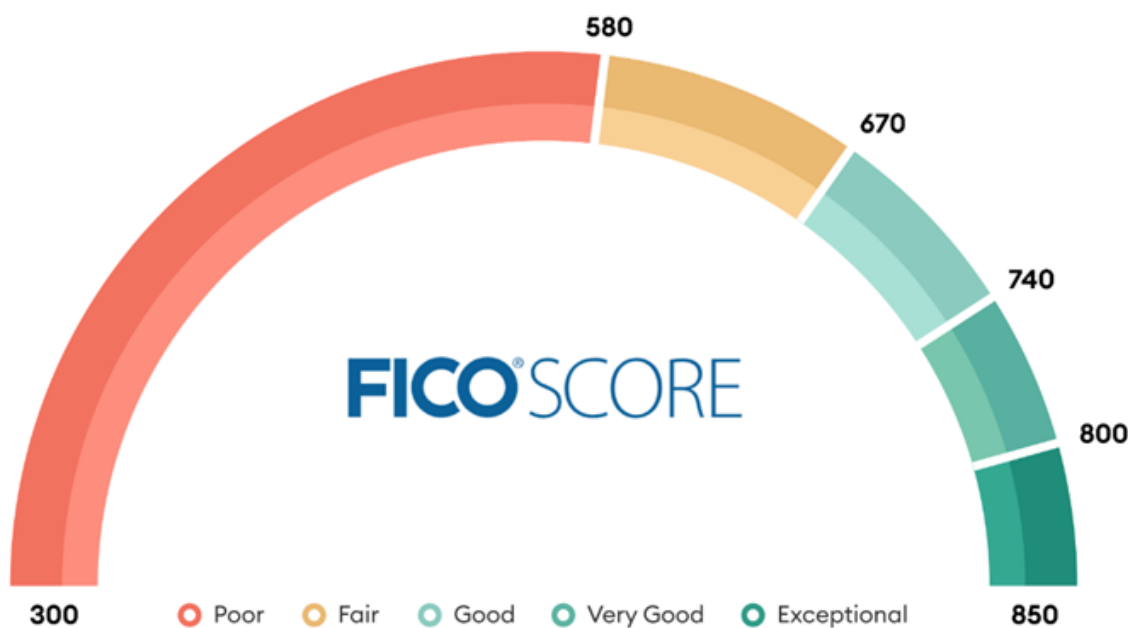
Biuro Informacji Kredytowej nie ukrywa jakie cechy są preferowane przez banki oraz na co należy zwrócić uwagę, aby podnieść kwotę maksymalnego możliwego do uzyskania kredytu (bik.pl, 2022):

- Dochody i forma zatrudnienia - umowa o pracę to zdecydowanie wariant, który wpływa dobrze na zdolność kredytową. W przypadku umowy o dzieło czy samozatrudnienia bank ma mniejszą pewność co do stabilności zarobków. Mniej oczywista jest również kwestia dochodów mikroprzedsiębiorców;
- Obciążenia wydatkami - Koszty stałe, kredyty, pożyczki, karty kredytowe - im mniej z nich wnioskodawca posiada tym potencjalnie wyższa miesięczna rata jaką może obsłużyć;

- Warunki wnioskowanego kredytu:
 - Czas - wydłużenie okresu kredytowania może obniżyć wysokość miesięcznej raty, co ma bezpośredni wpływ na zdolność kredytową;
 - Wspólny kredyt - rodzice, partner, partnerka, rodzeństwo to potencjalnie dobrzy kandydaci do wspólnego kredytu. Takie rozwiązanie nie tylko poprawia zdolność kredytową, ale jednocześnie daje dodatkowe zabezpieczenie dla banku;
 - Równe raty - lepiej wybrać kredyt o równych ratach kapitałowo-odsetkowych, aby zwiększyć swoje szanse na pozytywną decyzję kredytową;
- Historia kredytowa w BIK- z Raportu BIK można dowiedzieć się jakie informacje na temat wnioskodawcy dotychczas przekazywały do BIK banki, SKOK-i i firmy pożyczkowe.

1.2 Wyznaczanie oceny punktowej

W Stanach Zjednoczonych score kredytowy zawiera się w granicach od 300 do 850 punktów, gdzie w zależności od instytucji za odpowiednio dobry wynik traktuje się wartości powyżej 661-670 punktów(experian.com, 2021). Przedziały ocen deklarowane przez FICO (instytucję opisano szerzej w podrozdziale 1.3) przedstawiono na rysunku 2.




Rysunek 2: Diagram oceny wiarygodności kredytowej według firmy FICO(forbes.com, 2021)

W sieci można znaleźć darmowe kalkulatory score’u kredytowego i choć nie należy bezkrytycznie ufać ich kalkulacjom, jako że nie znamy dokładnych mechanizmów oraz zbiorów danych na podstawie których zbudowano dany model, to niektóre z nich potrafią zwrócić wyniki, które z pewną dozą niepewności możemy przyjąć jako prawdopodobne wartości wyliczane w profesjonalnych instytucjach. Jednym z takich narzędzi jest kalkulator na stronie CalcXML, który zwraca wynik w skali FICO. Osoba zainteresowana obliczeniem swojej indywidualnej oceny powinna przygotować kilka podstawowych informacji na swój temat(calcxml.com, 2023):

- Czy Wnioskujący posiadał pożyczkę lub kartę kredytową przez okres dłuższy niż 6 miesięcy;
- Ile lat temu Wnioskujący po raz pierwszy wziął pożyczkę lub posiadał kartę kredytową;
- Które z poniższych zobowiązań Wnioskujący posiada lub posiadał:
 - Kredyt hipoteczny;
 - Karta kredytowa;
 - Pożyczka na samochód/Pożyczka studencka/Inna pożyczka.
- Suma limitów ze wszystkich posiadanych przez Wnioskującego kart kredytowych;
- Czy Wnioskującego kiedykolwiek dotyczyło którekolwiek z poniższych ”negatywnych zdarzeń”:
 - Bankructwo;
 - Interwencja komornika/przejęcie własności;
 - Problemy podatkowe;
 - Inne ”negatywne zdarzenie”.
- Kiedy miało miejsce ostatnie ”negatywne zdarzenie”, które dotyczyło Wnioskującego.

Przykładowa kalkulacja, dla osoby posiadającej kartę kredytową z limitem 2000 dolarów od ponad roku, nie posiadającej dodatkowych zobowiązań, która w ostatnich dwunastu miesiącach nie wysyłała wniosków o kredyt, nie spóźniała się z spłatą zobowiązań oraz



Input And Assumptions

Have you had a credit card or loan for at least 6 months?

Yes

How many years ago did you get your first credit card or loan? (0 to 120)

1

Checkmark each type of credit account or loan that you have on your credit report, whether open or closed.

☐ Mortgage
☒ Credit Card
☐ Auto Loan
☐ Student Loan
☐ Other Loan
☐ Consumer Finance Account

How many times have you applied for credit in the last year?

0 times

When did you last miss a payment on any of your credit accounts?

Never

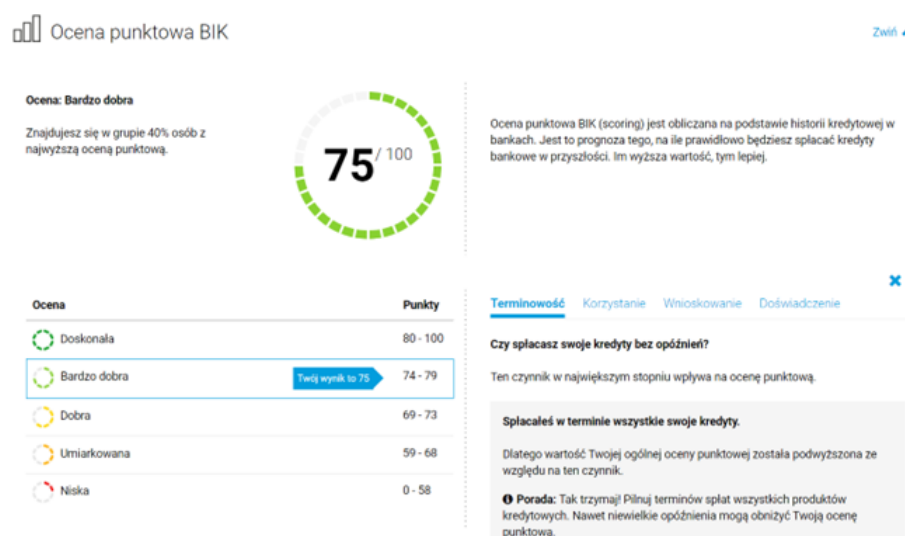
What is your total credit limit? (Add up the credit limits on all your credit card accounts.) (\$)

2,000

Rysunek 3: Fragment kalkulacji ze strony calcxml.com (calcxml.com, 2023)

nie brała udziału w żadnym z "negatywnych zdarzeń", której fragment przedstawiono na rysunku 3, daje wynik od 740 do 790 punktów, co oznacza bardzo dobry score kredytowy.

Od 21 grudnia 2017 roku BIK generuje ocenę punktową w nieco inny sposób. Dotychczas score zawierał się w zakresie od 192 do 631 punktów, co wizualizowane było za pomocą gwiazdek, gdzie im więcej gwiazdek, tym wyższa szansa na otrzymanie kredytu. Aby uczynić reprezentację graficzną bardziej czytelną dla osób fizycznych, ocenę przekształca się do postaci od 1 do 100 (totalmoney.pl, 2020) Zostało to zilustrowane w postaci wykresu kołowego na rysunku 4.



Rysunek 4: Wizualizacja oceny punktovej w BIK (źródło opracowanie własne)

Przed zmianą z grudnia 2017, Biuro Informacji Kredytowej do wyliczania oceny punktowej wykorzystywało tzw. Model II generacji o nazwie BIKSco CreditRisk. Dla tego algorytmu zakres możliwych do uzyskania rezultatów zawierał się pomiędzy 192, a 631 punktów i nie był w żaden sposób przekształcany. W najnowszych raportach wykorzystywany jest model III generacji o nazwie BIKSco CreditRisk 3, a punktacja zawiera się od 98 do 711 punktów. Jednakże otrzymany wynik ulega przekształceniu i tak jak jest to widoczne na rysunku 4, przedstawiony jest w zakresie od 1 do 100, co może mylnie wskazywać, że jest to procent maksymalnej, możliwej do uzyskania punktacji (scoringexpert.pl, 2018).

Bazując na cytowanym źródle, wynik wyliczonej punktacji podlega procesowi normalizacji według równania 1:

Równanie 1. Wzór na przekształcenie oceny wyliczonej z modelu BIKSco CreditRisk 3 do przedziału znormalizowanego 1-100 (experian.com, 2021)

$$S_{new} = \frac{S - min}{max - min} * (new_{max} - new_{min}) + new_{min}$$

gdzie:

- S_{new} - ocena punktowa z raportu BIK;
- S - oryginalna ocena punktowa z modelu BIKSco CreditRisk 3;
- min - minimalna wartość punktów z modelu BIKSco CreditRisk 3 (wynosi 98);
- max - maksymalna wartość punktów z modelu BIKSco CreditRisk 3 (wynosi 711);
- new_{max} - maksymalna wartość punktów z nowego zakresu (wynosi 100);
- new_{min} - minimalna wartość punktów z nowego zakresu (wynosi 1).

Na podstawie wartości widocznej dla osoby fizycznej (punktacji z przedziału 1-100), można uzyskać wynik wyliczony z algorytmu na podstawie równania 2:

Równanie 2. Wzór na wyliczenie wartości otrzymanej z modelu BIKSco CreditRisk 3 na podstawie wartości uzyskanej z Biura Informacji Kredytowej (experian.com, 2021)

$$S = \frac{613 * S_{new} + 9089}{99}$$

Biuro Informacji Kredytowej proponuje swoją interpretację znormalizowanej oceny punktowej, przedstawioną na rysunku 4, według której potencjalny kredytobiorca może ocenić swoje aktualne szanse na otrzymanie pożyczki.

Na podstawie cytowanej tabeli 1, można zinterpretować również ocenę nieznormalizowaną.

Ocena punktowa BIK	Słowna ocena	Komentarz
550+ (74+)	Bardzo dobra	Twój „scoring BIK” przewyższa średni „scoring BIK” Polaków. W ocenie banków Twoja wiarygodność kredytowa powinna być bardzo wysoka
500-549 (66-73)	Dobra	Twój „scoring BIK” oscyluje blisko średniego „scoringu BIK” Polaków. Banki zapewne ocenią Cię jako rzetelnego kredytobiorcę
400-499 (52-65)	Przeciętna	Twój „scoring BIK” jest poniżej średniego „scoringu BIK” Polaków. Tylko część banków oceni Twoją wiarygodność kredytową jako wystarczającą do uzyskania kredytu
poniżej 400 (poniżej 52)	Słaba	Twój „scoring BIK” jest znacznie poniżej średniego „scoringu BIK” Polaków. Taki scoring wskazuje, że jesteś ryzykownym kredytobiorcą dla banków. Możesz więc mieć problem z uzyskaniem kredytu, jeśli bank oprze się na „scoringu BIK” przy ocenie Twojego ryzyka kredytowego

Tabela 1: Interpretacja oceny punktowej BIK(scoringexpert.pl, 2017)

Za jedną z najbardziej znanych metod Credit Scoring’u jest uważana przytaczana na łamach tej pracy amerykańska metoda FICO (Fair, Isaac and Company). Zdefiniowana w 1989 roku, opiera się na 5 czynnikach(pl.economy pedia.com, 2021):

- Historia płatności (35 procent punktów) – na bazie dotychczasowych zobowiązań ocenia się, czy dana osoba wywiązuje się z nich na czas;
- Wykorzystanie kredytu (30 procent) – jeśli potencjalny klient instytucji finansowej dotychczas wykorzystywał niewielki procent dostępnych limitów kredytowych (np. limit na karcie kredytowej), ma on większe szanse na wyższą ocenę;
- Długość historii kredytowej (15 procent) – jeśli wnioskodawca jest doświadczonym

pożyczkobiorcą i przez długi czas poprawnie wywiązuje się z zobowiązań otrzymuje wyższą notę w tabeli punktowej;

- Nowe kredyty (10 procent) – złożenie wielu wniosków kredytowych przez osobę poszukującą kredytu, może wzbudzać wątpliwości instytucji przed udzieleniem pożyczki;
- Rodzaje wykorzystanego kredytu (10 procent) – dla banków mile widziane jest doświadczenie wnioskującego w zarządzaniu różnymi rodzajami kredytów (karta kredytowa, kredyt hipoteczny, pożyczka gotówkowa itp.).

Rozwój Credit Scoringu jest jednym z powodów, dla których rynek kredytów konsumenckich w Stanach Zjednoczonych w latach 90. XX w. eksplodował. Kredytodawcy czuli się bardziej pewni w udzielaniu pożyczek szerszym grupom ludzi, ponieważ mieli dokładniejsze narzędzie do pomiaru ryzyka. Scoring kredytowy pozwolił im również na szybsze podejmowanie decyzji, umożliwiając rozpatrzenie większej liczby wniosków, czego rezultatem był bezprecedensowy wzrost ilości dostępnych kredytów konsumenckich (Weston, 2012). Banki bardzo skrupulatnie podchodzą do operowania swoimi pieniędzmi, dokładnie „prześwietlając” swoich klientów pod kątem wypłacalności. Polskie instytucje finansowe często posiłkują się opinią BIK, jednakże równie chętnie stosują także własne algorytmy oceny, których szczegółów zwykle szerzej nie udostępniają.

1.3 Score kredytowy w erze Big Data

Historia rozwoju modeli scoringowych sięga tak daleko, jak historia pożyczania i spłacania. Widać to w szczególności w potrzebie ustalenia odpowiedniej stopy procentowej, uwzględniającej ryzyko braku możliwości odzyskania pożyczonych pieniędzy. Wraz z nadejściem współczesnej ery statystyki w XX. wieku rozpoczęto opracowywanie technik oceny prawdopodobieństwa niewykonania przez kredytobiorcę zobowiązania do zwrotu środków. Pod uwagę brano podobieństwo przypisanych cech do tych, którymi charakteryzują się osoby niewywiązujące się ze zobowiązań w przeszłości. Tego typu metody statystyczne są obecnie stosowane powszechnie przez praktycznie wszystkie banki, a również znaczną część pozostałych instytucji finansowych (prawniczydotblog.wordpress.com, 2019).

Dziedzina Credit Scoringu nie należy do odkryć bieżącej dekady, a jej początki można upatrywać w połowie XX wieku, wraz z utworzeniem w 1956 roku wspominatej już wcześniej firmy FICO. Przedsiębiorstwo założyli inżynier Bill Fair oraz matematyk Earl Judson Isaac (Wikipedia, 2022). Firma zajmowała się budową systemów scoringowych, jednakże przez długi czas stosowano zapis papierowy (StatSoft, 2010). Początkowo zajmowano się głównie procesami weryfikacji wniosków kredytowych w bankach, stosując proste tabele punktowe i głównie opierając się na tzw. metodzie eksperckiej (Thonabauer & Nosslinger, 2004). Sposób musiał być na tyle łatwy i intuicyjny, aby dawać możliwość obiektywnej oceny zdolności potencjalnego kredytobiorcy do wywiązania się z zaciągniętego zobowiązania kredytowego również mniej doświadczonym i niewykwalifikowanym pracownikom banku (Thomas, Edelman, & Crook, 2002). Jednym z pierwszych i najważniejszych osiągnięć Fair Isaac & Company było utworzenie pierwszego, wykorzystywanego komercyjnie, systemu scoringowego (Poon, 2007).

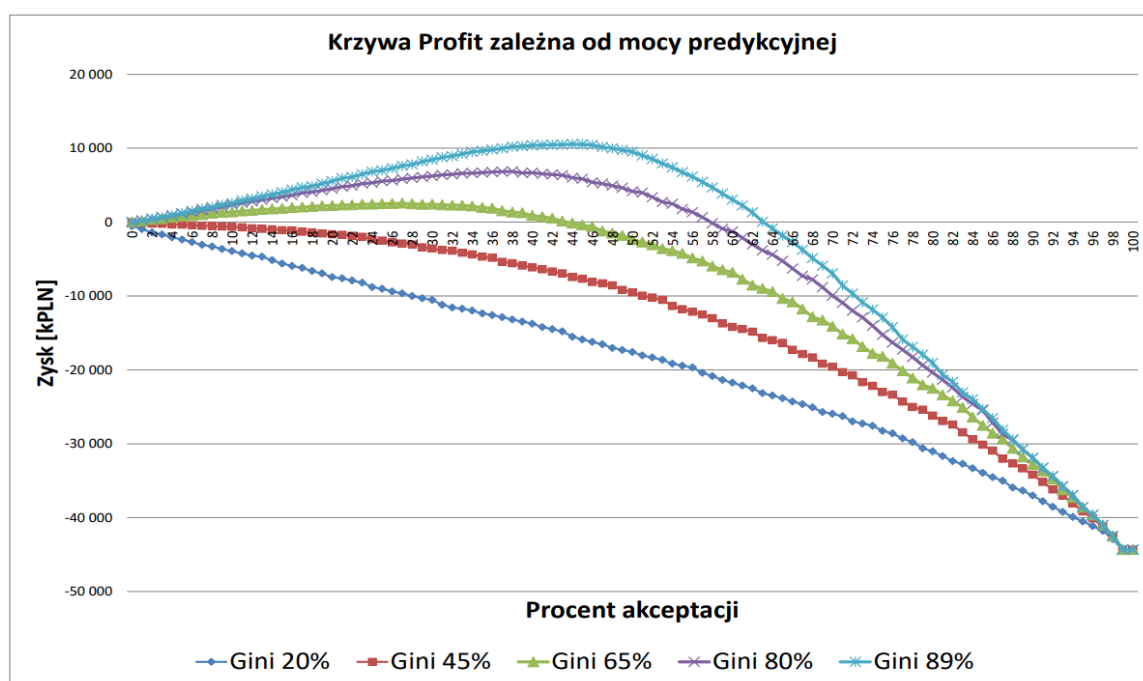
Jednakże zanim FICO rozpoczęło swoją działalność, ludzkość zebrała już pierwsze doświadczenia ze złymi kredytobiorcami, a co za tym idzie zaczęto zastanawiać się nad ich właściwym zidentyfikowaniem zanim zostanie im udzielona pożyczka. Rok 1826 przyjmuje się za początek wymiany informacji między wierzycielami, co miało na celu podniesienie jakości "filtracji" właściwych klientów, natomiast w 1899 roku w Atlancie rozpoczęto zbieranie danych na większą skalę. Po powstaniu Fair Isaac & Company rozwój w dziedzinie udzielania kredytów znacząco przyspieszył, co skutkowało powstawaniem konkurencyjnych dla FICO firm (wśród nich warto wymienić chociażby Experian czy Equifax), ale też było bodźcem do podwyższenia jakości usług (ecomparemo.com, 2020). Zwieńczenie i zarazem wykładnik wzrostu znaczenia score'u kredytowego stanowiło wejście Fair, Isaac and Company na giełdę w Nowym Jorku w lipcu 1987 roku (fico.com, 2023).

Wraz z rozwojem technik informatycznych rozpoczęto wdrażanie bardziej zautomatyzowanych procesów scoringowych. Najpowszechniej do tego celu wykorzystywano model regresji logistycznej, jednakże dziś stosuje się szeroką gamę różnych metod predykcyjnych tj. sieci neuronowe, lasy losowe czy drzewa decyzyjne (Przanowski, 2014). Analitycy i inżynierowie danych w dzisiejszych czasach wykorzystują potencjał płynący z zastosowania skomputeryzowanych metod predykcji zdarzeń.

Jednym z podstawowych podejść w modelowaniu credit scoring jest wyliczenie prawdo-

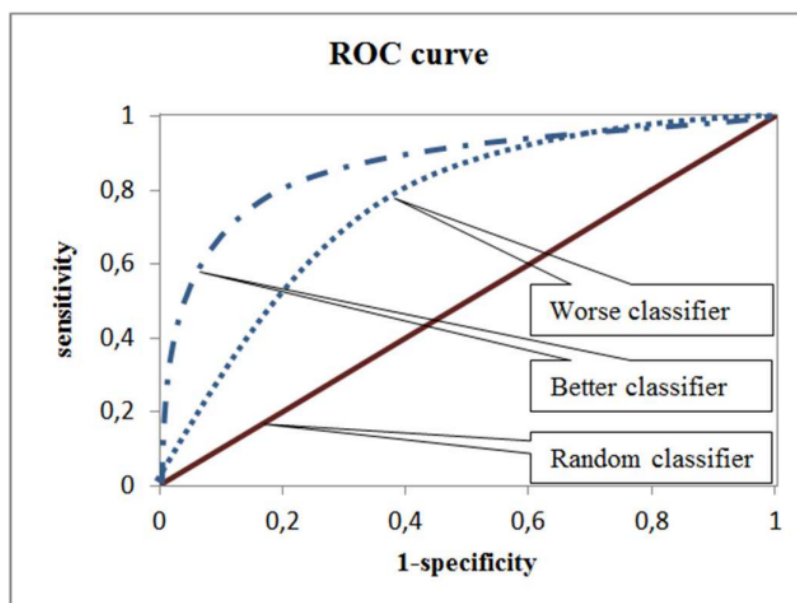
podobieństwa wystąpienia zdarzenia znanego pod nazwą "default". Finalnie, bazując na zebranych danych, zarówno tych opisujących klientów z przeszłości, jak również dotyczących wnioskodawcy, model powinien obliczyć procentową szansę na to, że wnioskodawca "wpadnie w default", gdzie np. zmienna default12 wskazuje, iż dłużnik w ciągu 12 miesięcy od zaciągnięcia długu spóźniał się ze spłatą raty o co najmniej 90 dni (Przanowski, 2015). Pozornie logika nakazywałaby odrzucać wnioski, dla których prawdopodobieństwo problemów ze spłatą wynosi więcej niż 50 %, jednakże nie jest to tak oczywiste, jak może się na pierwszy rzut oka wydawać. Celem zobrazowania tego zagadnienia warto jest omówić i zrozumieć działanie tzw. Krzywej Profit (rysunek 5).

Na Krzywej Profit wizualizuje się wartość zysku w zależności od procentu zaakcepto-



Rysunek 5: Krzywa Profit zależna od mocy predykcyjnej (Przanowski, 2023)

wanych wniosków dla różnych modeli predykcyjnych, gdzie każdy z nich opisuje się współczynnikiem Giniego. Model buduje się "trenując go" na zbiorze treningowym, złożonym z danych historycznych, gdzie algorytm wychwytuje zależności cechujące klientów spłacających zobowiązania bez opóźnień, a także uczy się odróżniać niewiarygodnych wnioskodawców. Następnie na zbiorze testowym porównuje się wnioski prognozowane przez model z sytuacją jaką w rzeczywistości miała miejsce i na tej podstawie buduje się tzw. krzywą ROC - przykładowa wizualizacja na rysunku 6.



Rysunek 6: Krzywa ROC i jej możliwe warianty (Gajowniczek et al., 2014)

Do skonstruowania krzywej ROC niezbędne jest wyliczenie czterech parametrów, składających się na tzw. macierz pomyłek (Fawcett, 2005):

- False Positive (FP) - model wskazał, że dana osoba wpadnie w default, mimo iż w rzeczywistości sumiennie spłacała kredyt;
- False Negative (FN) - model wskazał, że dana osoba nie wpadnie w default, mimo iż w rzeczywistości miała kłopoty ze spłatą;
- True Positive (TP) - model wskazał, że dana osoba wpadnie w default, co okazało się zgodne z rzeczywistością;
- True Negative (TN) - model wskazał, że dana osoba nie wpadnie w default, co okazało się zgodne z rzeczywistością.

Na podstawie powyższych wartości oblicza się wielkości takie jak czułość, swoistość itp., które pozwalają na zwizualizowanie jakości modelu na krzywej ROC. Jednakże do tego celu należy wyliczyć wiele punktów, a dokonuje się tego poprzez badanie powyżej opisanych właściwości w zależności od przyjętego punktu odcięcia (Fawcett, 2005).

W procesie predykcyjnym wyznaczane są prawdopodobieństwa wejścia w default, a decyzja o tym czy dana osoba zostanie zaklasyfikowana jako wystarczająco wiarygodna zależy od przyjętego poziomu akceptacji. Jeśli dla pewnego wnioskodawcy prawdopodobieństwo default zostało wyznaczone na 35%, a próg odcięcia założono na poziomie 40%,

kredyt zostanie udzielony. Jednakże w przypadku zdefiniowania punktu podziału równego 30%, wniosek zostanie odrzucony. W celu stworzenia wykresu jakości modelu należy wyliczyć wartości macierzy pomyłek dla wielu progów odcięcia (Fawcett, 2005). Po wykonaniu tych operacji możemy wyznaczyć wartość współczynnika Giniego dla rozpatrywanego algorytmu, który wynosi dwukrotność pola pod krzywą ROC, a nad krzywą klasyfikatora losowego (na rysunku 6 oznaczona jako "Random classifier").

Wracając do Krzywej Profit (rysunek 5), łatwo zauważyć, że im skuteczniejszy model (im wyższy współczynnik Giniego), tym wykres jest bardziej wybrzuszony, co oznacza możliwość osiągnięcia wyższych zysków. Jako że straty powodowane przez jednego dłużnika potrafią przewyższyć profity uzyskiwane przez bank we współpracy z wieloma wiarygodnymi klientami, bardzo istotny jest dobór odpowiedniego progu akceptacji. Przytoczony wykres, dla najlepszego z modeli, sugeruje wybór progu odcięcia pomiędzy 40%, a 46% (Przanowski, 2023).

Klasyczne karty punktowe cechują się prostotą w interpretacji wyników, a wykorzy-

Attribute	Variable	Partial Score
<20	Age	10
20>= and <34		20
35>=		30
Bad	Payment history	10
Not good		25
Good		40

Tabela 2: Klasyczna karta oceny punktowej (Przanowski, 2023)

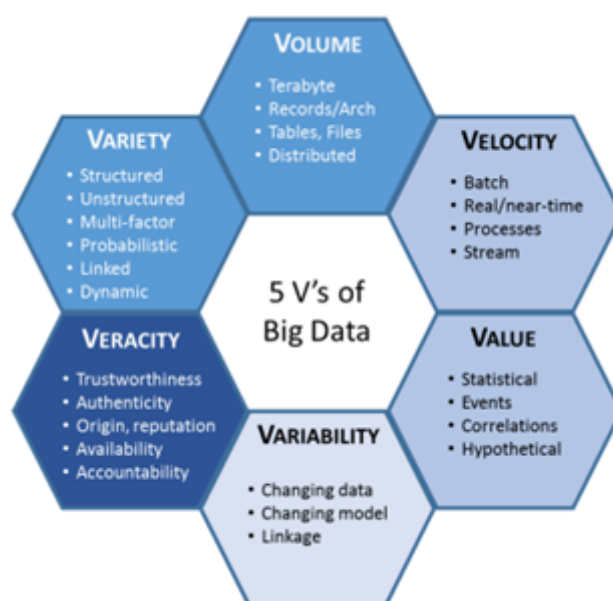
stanie takich jak tabela 2 nie wymaga wiedzy analitycznej wymaganej w stosowanych współcześnie metodach komputerowych. Trudno zatem oprzeć się wrażeniu, iż klasyczny Credit Scoring usuwa się w cień na korzyść narzędzi z dziedziny Big Data, choć w rzeczywistości staje się jedną z poddziedzin szerokiej tematyki „dużych danych” (Przanowski, 2014). Lecz czym właściwie jest Big Data?

Pojęcie to powinno iść w parze z rozbudowanymi systemami informatycznymi, które dają możliwość przetwarzania dużych danych. Często spotykanym jest tzw. 5V Big Data (zwizualizowane na rysunku 7 (researchgate.net, 2017)), dające ogólne spojrzenie na podstawowe cechy tejże dziedziny (techtarg.com, 2021):

- Volume (ang. wolumen, wielkość) – jeśli mamy do czynienia z subiektywnie dużą

ilością danych (co dokładnie oznacza „duża ilość” nie zostało precyzyjnie zdefiniowane), możemy nazwać ich przetwarzanie jako Big Data;

- Velocity (ang. szybkość) – ze względu na ich ilość, dane muszą być odpowiednio szybko procesowane, najczęściej w czasie rzeczywistym;
- Variety (ang. różnorodność) – technologia musi radzić sobie z operowaniem na danych zróżnicowanego, nie koniecznie uporządkowanego typu;
- Veracity (ang. wiarygodność) – dane najczęściej nie są wysokiej jakości, a ich braki czy też błędne informacje w nich zawarte stawiają wymóg odpowiedniej odporności na tego rodzaju zaburzenia;
- Value (ang. wartość) – kluczowa przy przetwarzaniu danych jest możliwość uzyskania na ich bazie istotnych informacji, które mogą wnieść pewną wartość dla przedsiębiorstwa, dokonującemu lub zlecającemu wykonanie takich analiz.



Rysunek 7: Schemat 5V Big Data(researchgate.net, 2017)

Początkowo Credit Scoring specjalizował się głównie we wspomaganiu procesów decyzyjnych w bankach, a narzędzia Big Data stosowane były w globalnych firmach świadczących usługi w świecie wirtualnym tj. Google, Amazon czy Facebook. Z kolei w Polsce zarządzaniem dużymi danymi na poważnie zainteresowały się jako pierwsze Onet czy portal Nasza Klasa(Przanowski, 2014). Mimo zainteresowania różnymi branżami, Big Data i Credit Scoring poruszają podobne problemy ze strony merytorycznej, gdzie głównym

i najpoważniejszym problemem zawsze był kluczowy element ich funkcjonowania – dane.

Modele Credit Scoring służą do prognozowania zjawisk na podstawie dotychczas zaobserwowanej i zebranej historii danych. Proces spłacania kredytów najczęściej trwa wiele lat, zatem potrzeba dużo czasu, aby zebrać dostatecznie reprezentatywną pulę informacji rzeczywistych, którą następnie można wykorzystać do sprawdzenia użyteczności i poprawności skonstruowanego modelu (Przanowski, 2014).

W przypadku danych bankowych, sytuacja jest jeszcze trudniejsza z uwagi na wrażliwość informacji. Skutkuje to koniecznością występowania do instytucji finansowych z oficjalnymi podaniami, a otrzymane dane często są zafałszowane i zanonimizowane, co zwykle uniemożliwia ich zinterpretowanie. Znacząco utrudnia to tworzenie odpowiednio wiarygodnych modeli scoringowych, na co analitycy odpowiadają tworzeniem własnych, symulowanych danych (Przanowski, 2014).

Dziedzina Big Data nie jest bez wad i mimo jej niekwestionowanej przydatności, bez trudu można wymienić niepowodzenia i wyzwania jakie napotyka, gdzie najistotniejsze z nich to (Przanowski, 2023):

- Dane często nie są zbierane, a jeśli ktoś już je magazynuje, nie zapewnia odpowiedniej ich przydatności i interpretowalności;
- Problem jakości danych;
- Liczne braki danych;
- Brak publicznych danych, dostępnych i przykładowych;
- Brak inwestycji w przygotowanie i wykształcenie inżyniera danych.

1.4 Metody Uczenia Maszynowego wykorzystywane w dziedzinie Credit Scoring

Dotychczas w bankowości nie stosowano powszechnie niektórych technik Machine Learningu (ML) do zarządzania ryzykiem, a geneza takiego postępowania była zrozumiała – modele są trudne w interpretacji, a ponadto generują popyt na wysoce wyspecjalizowanych pracowników. Z drugiej zaś strony, rynek konkurencyjny zmienia się - transformacja cyfrowa, czy też nowy model bankowości otwartej wywierają wpływ na praktyki zarządzania ryzykiem. W tym kontekście stosowanie technik Uczenia Maszynowego zapewnia

istotną przewagę, skracając czas podejmowania decyzji w procesach kredytowych oraz podnosząc ich skuteczność (crif.pl, 2018).

Techniki stosowane do opracowania kart scoringowych to np. dyskryminacja statystyczna i metody klasyfikacji. Należą do nich modele regresji liniowej (łatwo interpretowalne, oparte na metodzie minimalizacji sumy kwadratów reszt), analiza dyskryminacyjna (odmiana regresji stosowana do klasyfikacji), modele logitowe i probitowe (maksymalizacja prawdopodobieństwa zaobserwowania wartości), oraz tzw. modele eksperckie (np. proces analizy hierarchicznej – AHP)(Group, 2021).

Natomiast główne metody Uczenia Maszynowego, jakie stosuje się przy Credit Scoring’u należą do grupy technik klasyfikacji nadzorowanej(Bajek, 2011). W dalszej części pracy, celem dokładniejszego opisanie, skupiono się na dwóch, najpopularniejszych metodach stosowanych do budowy modeli scoringowych, jakimi są regresja logistyczna i drzewa decyzyjne, pomijając równie ciekawe, aczkolwiek rzadziej używane podejścia tj. sieci neuronowe, lasy losowe czy też SVM – Support Vector Machines (ang. Metoda Wektorów Nośnych)(Statsoft, 2010).

Regresja logistyczna (LR) jest metodą statystyczną umożliwiającą ocenę wpływu wielu cech - tzw. zmiennych objaśniających - na szanse zajścia zdarzenia, np. zachorowania na pewną chorobę, czy też spłaty kredytu w planowanym terminie(Deryło, 2021). Model LR jest szczególnym przypadkiem uogólnionego modelu liniowego. Znajduje zastosowanie, gdy zmienna zależna jest dychotomiczna, to znaczy przyjmuje tylko dwie wartości takie jak np. sukces lub porażka, zwykle reprezentowane jako cyfry 1 i 0(statystyka.az.pl, 2021).

Starając się jak najkrócej scharakteryzować to podejście, należy podkreślić, że głównym celem modeli regresji logistycznej jest znalezienie najlepszych współczynników(tzw. wag), które minimalizują błąd pomiędzy przewidywanym prawdopodobieństwem, a obserwowanym wynikiem. Realizowane jest to za pomocą algorytmu optymalizacji, takiego jak opadanie gradientu, celem dostosowania współczynników, aż model będzie odpowiednio pasował do danych, na których jest uczony, co następnie walidowane jest na zbiorze testowym(newsblog.pl, 2022). Wartości zmiennej dychotomicznej możemy przekształcić w postać prawdopodobieństwa wystąpienia danego zdarzenia, które przyjmuje wartości pomiędzy 0 lub 1. Gdy zastosuje się transformację logit możliwe jest zlinearyzowanie modelu LR i przedstawienie go w postaci regresji liniowej(naukowiec.org, 2014).

Regresja logistyczna jest najczęściej wykorzystywaną techniką modelowania scoringowego. Jej największą zaletą jest stabilność w czasie, co czyni ją względnie odporną na zaburzone dane i może sprawiać wrażenie, że metoda jest odporna na ataki. Jej zaleta może być również interpretowana jako wada, ponieważ model może powodować nieoptymalność dopasowania do koniunktury rynkowej (Karolak, 2014).

Jako przykład implementacji modelu regresji logistycznej można przedstawić badanie wpływu wysokości dochodów na fakt palenia papierosów przez badaną próbę osób. Przyjmijmy, że zmienna niezależna, jaką jest w tym przypadku ilość dochodu, jest zmienną ilościową, a palenie papierosów określamy binarnie, gdzie 1 oznacza, że osoba pali papierosy, a 0 zdarzenie przeciwne. Jedna z takich analiz wskazała, że poziom dochodów jest istotny statystycznie, a osoby zarabiające więcej mają wyższą skłonność do palenia wyrobu tytoniowego (naukowiec.org, 2014).

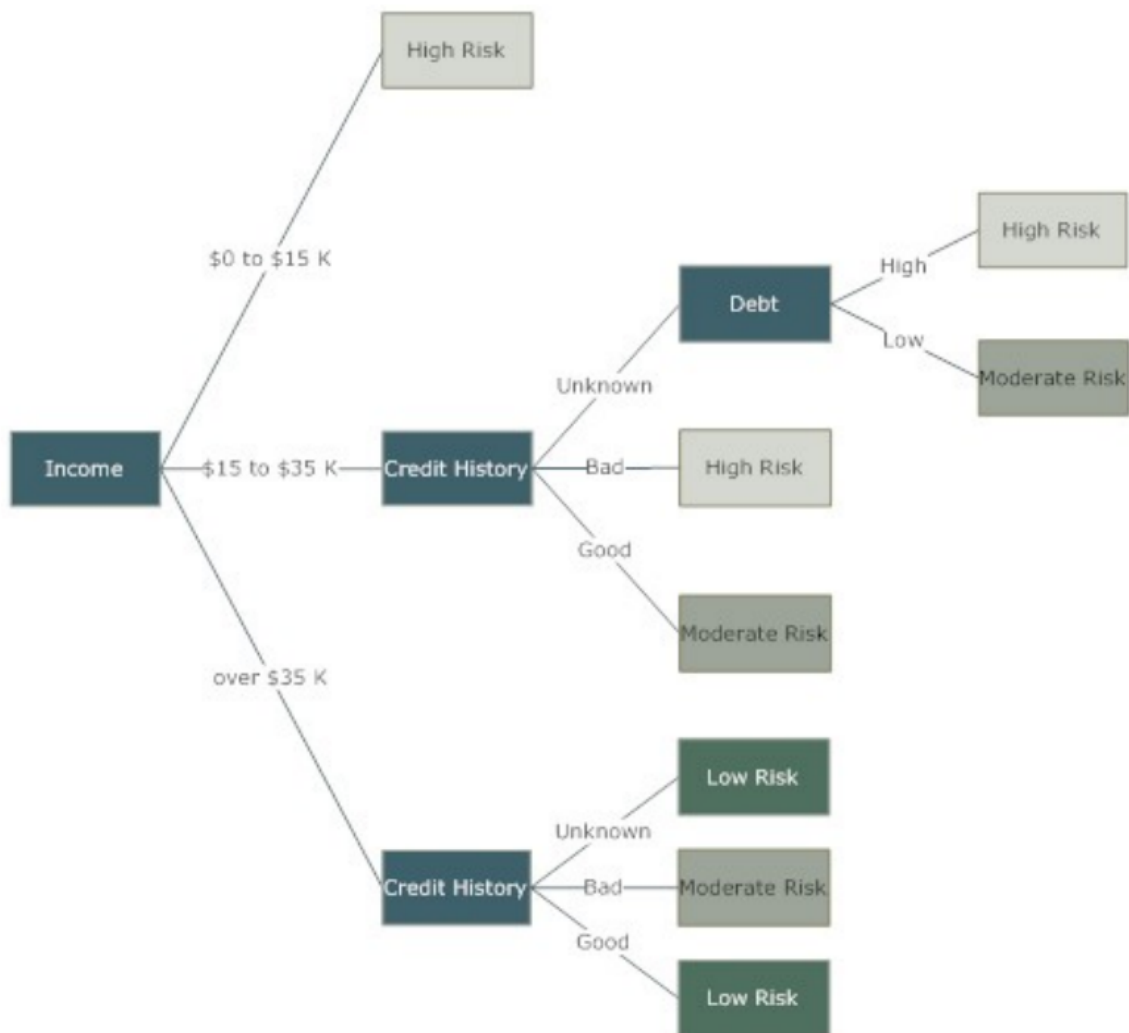
Drzewa decyzyjne (DT) są modelami o strukturze drzewiastej, podejmującymi decyzje na podstawie wartości poszczególnych cech. Można ich używać zarówno do klasyfikacji binarnej (służy do przewidywania, do której z dwóch klas/kategorii należy wystąpienie danych), jak i wieloklasowej (odnosi się do zadań klasyfikacyjnych, które mają więcej niż dwie etykiety klas). Pośród zalet drzew decyzyjnych należy wymienić prostotę w budowie oraz łatwość w interpretacji, dzięki czemu można dokładnie prześledzić cały proces podejmowania decyzji i jasno określić konkretne kryteria, na bazie których zrealizowano dany podział. Ich nauka jest szybka, a poza tym radzą sobie z zarówno liczbowymi jak i kategorycznymi typami danych (newsblog.pl, 2022). Są również bardzo skuteczne w przetwarzaniu znacznych ilości danych i nie wymagają ponadprzeciętnych mocy obliczeniowych (Bujak, 2008).

Nie wolno również zapomnieć o zagrożeniu jakie ze sobą niosą. Przy ich konstrukcji należy uważać na parametry, ponieważ nadmierna głębokość drzewa, czy też zbyt wiele utworzonych gałęzi może prowadzić do przeuczenia – nadmiernego dopasowania do danych treningowych – co spowoduje nieefektywne oszacowania wartości prawdopodobieństwa zajścia zdarzenia w środowisku produkcyjnym (newsblog.pl, 2022).

Drzewa decyzyjne mają ustalony porządek(Bujak, 2008):

- korzeń odpowiada wszystkim możliwym decyzjom;
- każdy wewnętrzny węzeł odpowiada pewnej decyzji, którą możemy podjąć;
- liściom odpowiadają cele.

Algorytm przedstawiono również na przykładzie prostej decyzji kredytowej na rysunku 8:



Rysunek 8: Drzewo decyzyjne zastosowane do kategoryzacji potencjalnych kredytobiorców pod kątem ryzyka kredytowego (Bujak, 2008)

Rozwój Uczenia Maszynowego na początku XXI wieku, może dawać nadzieję na automatyzację w wielu branżach. Jesteśmy coraz bliżej powszechnego wykorzystania pojazdów autonomicznych, a od wielu lat modele ML są implementowane w dynamicznych

start-up'ach, czy też w badaniach na uczelniach ekonomiczno-technicznych, stale udowadniając jak wszechstronne mogą być zastosowania systemów uczących się. Ze względu na ich mnogość i zróżnicowanie, poprzez odpowiedni dobór metod, analitycy podnoszą skuteczność stosowanych rozwiązań, dając satysfakcję z wysokiego poziomu poprawnych predykcji uzyskanych z modeli. Stajemy się coraz bardziej zależni od wyników otrzymywanych z analiz ML, czego przykładem są modele Credit Scoring. Dziś to od maszyny zależy, czy wnioskodawca otrzyma kredyt i zrealizuje marzenie o własnym mieszkaniu, czy nowym telewizorze. Jednakże wraz ze wzrostem roli maszyn w naszym codziennym życiu, pojawia się pytanie – czy są one bezpieczne? Temat odporności systemów Uczenia Maszynowego został poruszony w kolejnym rozdziale.

Credit Scoring jest jednym z kluczowych narzędzi wykorzystywanych przez instytucje finansowe. W czasach gdy powszechny konsumpcjonizm i rozpędzone gospodarki świata generują potrzebę kreacji pieniądza, kredyt jest jedną z pierwszych rzeczy przychodzących na myśl. Zarówno wielkie korporacje, jak i osoby prywatne od czasu do czasu potrzebują znaczącego zastrzyku gotówki np. na kupno mieszkania, wymarzony samochód czy dobrze prosperującą inwestycję. Banki dziesiątki lat temu zauważyły potencjał leżący w tej dziedzinie analizy, a rozwój technologii może jeszcze bardziej zwiększyć niezawodność systemów predykcyjnych. Należy jednak pamiętać, że nawet najnowocześniejsze modele uczenia maszynowego potrzebują odpowiedniej puli wiarygodnych danych, aby najpierw skutecznie nauczyć się na błędach sprzed lat, a następnie zapobiec złym kredytobiorcom w przyszłości.

2 Adversarial Machine Learning – Wrogie Uczenie Maszynowe

Modele Uczenia Maszynowego otworzyły zupełnie nowe możliwości w dziedzinie automatyzacji, a wizja wszechobecnej Sztucznej Inteligencji skutecznie rozpala wyobrażenia ludzi o świecie, w którym na porządku dziennym będziemy wykorzystywać roboty, posiadające własną świadomość. Jednakże należy pamiętać, że z wielką mocą wiąże się wielka odpowiedzialność, a wykorzystanie nowych możliwości w złym celu, może nieść ze sobą groźne skutki. W kolejnym podrozdziale pochyłono się nad problemem Wrogiego Uczenia Maszynowego, czyli potencjalnych ataków na modele ML.

2.1 Wprowadzenie do technik Uczenia Maszynowego

Karty punktowe, mimo że łatwe w interpretacji zarówno przez wnioskujących, jak i sprzedawców, nie są najbardziej optymalnym narzędziem do oceny wiarygodności kredytowej. Przez ostatnie kilkadziesiąt lat pojawiło się wiele nowych możliwości analizy, co jest związane z nieustającym rozwojem informatyzacji, a niektóre z nich zostały dopasowane do dziedziny Credit Scoring’u, dając lepszą efektywność predykcji oraz podnosząc zyski instytucji finansowych.

Uczenie Maszynowe, samouczenie się maszyn albo systemy uczące się, w języku angielskim tłumaczone jako Machine Learning (ML) jest dziedziną wchodzącą w skład nauk, zajmujących się Sztuczną Inteligencją (Artificial Intelligence – AI), Głównym jej celem jest tworzenie automatycznego systemu, który potrafi doskonalić się na bazie doświadczenia i nabywać na tej podstawie nową wiedzę. W uproszczeniu proces polega na znalezieniu wzorca w dostarczonych danych. Modele Uczenia Maszynowego powszechnie wykorzystywane są w wielu dziedzinach, w których zachodzi potrzeba predykcji pewnego zjawiska(gov.pl, 2021).

Zadania ML ograniczone są do wąskiego, specyficznego zakresu, w którym ma działać dany system. W przeciwieństwie do sztucznej inteligencji, proces uczenia maszynowego nie jest w stanie stworzyć czegoś nowego, a jedynie uzyskiwać najbardziej optymalne rozwiązania w zadanym problemie. Najpopularniejszymi aplikacjami wykorzystującymi możliwości Uczenia Maszynowego są wyszukiwarki online, algorytmy podpowiadające najciekawsze dla użytkowników materiały w mediach społecznościowych, rozpoznawanie obrazów czy filtrowanie spamu ze skrzynek e-mail(elektronikab2b.pl, 2021).

Machine Learning stało się popularne relatywnie niedawno, ale jego historia jest znacznie dłuższa. Najważniejszy czynnik w analizie danych, czyli właśnie dane, zaczęto zbierać już w czasach starożytnych, a były to informacje o ilości zgromadzonej żywności, czy też szczegóły dot. spisów powszechnych. Kiedy z biegiem lat danych przybywało, ich analiza stawała się trudniejsza, a przełomowym momentem okazała się siedemnastowieczna epidemia dżumy, dziesiątkująca mieszkańców Anglii, gdy wówczas zaczęto publikować pierwsze zbiory danych dotyczące zdrowia publicznego. Jednakże przetwarzanie dużych zbiorów było procesem żmudnym, a jeszcze w drugiej połowie XIX wieku zebranie i przeanalizowanie danych z przeprowadzonego w Stanach Zjednoczonych spisu powszechnego zajmowało nawet dziesięć lat(fotc.com, 2022).

W latach 50. XX w. Alan Turing, znany z udziału w rozszyfrowaniu Enigmy, niemieckiej maszyny szyfrującej, stwierdził, że „jeżeli maszyna będzie w stanie przekonać człowieka, że wcale nie jest maszyną, to będzie to świadczyć o osiągnięciu przez nią sztucznej inteligencji”(fotc.com, 2022). Z kolei Artur Samuel w latach 1952-1962 rozwijał program do szkolenia zawodników szachowych, jak również on, na konferencji w 1959 roku, po raz pierwszy użył pojęcia Uczenie Maszynowe jako „[...] dające komputerom możliwość „uczenia się” bez bycia konkretnie zaprogramowanym do danego zadania.”(Mamczur, 2019).

W roku 1957 Frank Rosenblatt opracował pierwszą komputerową sieć neuronową, która wczytując obrazy, generowała etykiety i kategoryzowała ilustracje(fotc.com, 2022). Kolejnym znanym systemem był Dendral z 1965 roku, który powstał na Uniwersytecie Stanforda z inicjatywy dwóch naukowców - Edwarda Feigenbauma oraz Joshuy Lederberga. Celem programu była automatyzacja analiz i identyfikacji molekuł związków organicznych nieznanymi ówczesnym chemikom(britannica.com, 2019).

Przez kolejne dekady było dość cicho w zakresie Uczenia Maszynowego, a do lat dwięćdziesiątych stosowano je głównie w prostych grach. Następnie ML było coraz częściej wykorzystywane przede wszystkim przez przeglądarki internetowe tj. Google czy Yahoo oraz wspomagało systemy anty-spamowe. Wraz z początkiem drugiej dekady XXI wieku o Uczeniu Maszynowym ponownie zrobiło się głośno, głównie dzięki rozwojowi sieci neuronowych(Mamczur, 2019). Badania prowadzone wiele lat temu pozwoliły na to, by wiedza o ML nie była tak egzotyczna, a przeciętny student kierunku związanego z branżą informatyczną potrafił podać co najmniej kilka jej zastosowań w życiu codziennym.

Jednakże aby modele oparte na algorytmach Machine Learning'u mogły być zastosowane w biznesie muszą być odpowiednio dostosowane do zagadnienia, a co za tym idzie – nauczone na bazie wprowadzonych danych. Uczenie maszynowe nie jest jednolitą technologią, a sposób jej działania zależy w dużej mierze od tego, z jakich algorytmów korzysta i jakimi danymi zostanie zasilona.

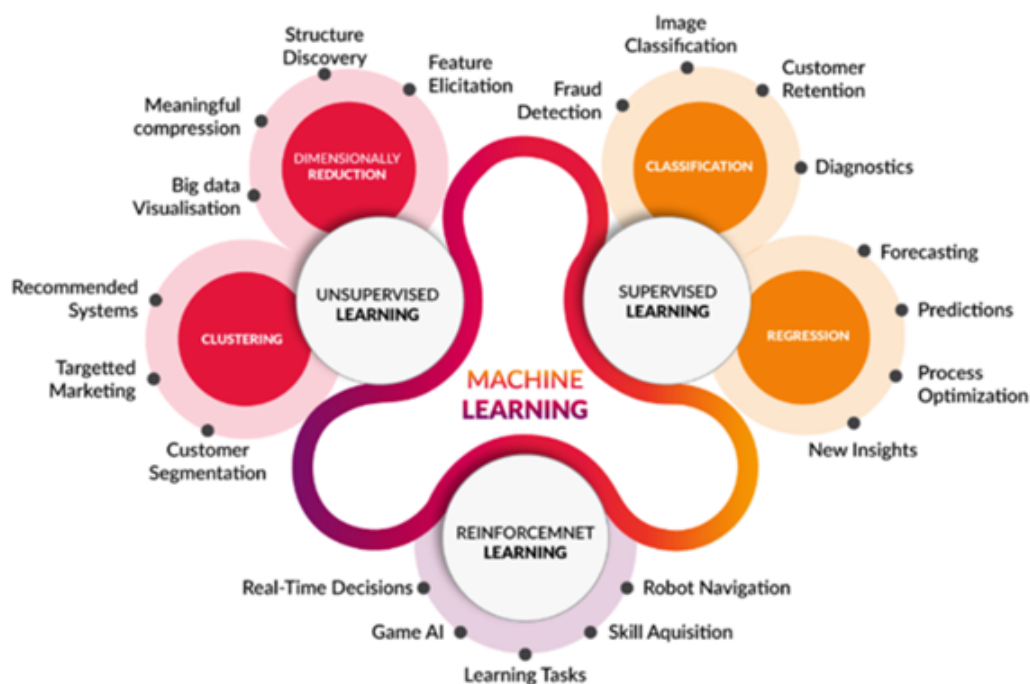
Eksperti SAS wskazują 4 podstawowe techniki uczenia maszynowego(sas.com, 2018):

- Supervised Learning (ang. Uczenie Nadzorowane) - maszyny uczą się na podstawie dostarczonych przykładów, a dane wejściowe są wykorzystywane do wyszukiwania zależności, służących do rozwiązania określonego problemu. Gdy uda się ustalić pewien wzorzec, jest on wykorzystywany w podobnych przypadkach. Do przykładowych zastosowań tej metody należą:
 - zarządzanie ryzykiem;
 - rozpoznawanie mowy, tekstu i obrazu;
 - segmentacja klientów.
- Semi - Supervised Learning (ang. Uczenie Częściowo Nadzorowane) - maszyna otrzymuje zarówno dane wejściowe oznaczone (zawierające odpowiadające im dane wyjściowe, konkretne przykłady), jak i nieoznaczone (wymagające przyporządkowania do danych wyjściowych, znalezienia odpowiedzi). Ten rodzaj uczenia wykorzystuje się w sytuacjach, gdy dana instytucja dysponuje zbyt dużą ilością danych lub gdy informacje cechują się wysokim zróżnicowaniem, które uniemożliwia przyporządkowanie odpowiedzi do każdej z nich. W takiej sytuacji system sam proponuje odpowiedzi i jest w stanie stworzyć ogólne wzorce. Metoda znajduje zastosowanie w:
 - rozpoznawaniu mowy;
 - rozpoznawaniu obrazów;
 - klasyfikacji stron internetowych.
- Unsupervised Learning (ang. Uczenie Nienadzorowane) - maszyna nie posiada „klucza odpowiedzi” i musi sama analizować dane, szukać wzorców i odnajdować relacje. Ten typ ML najbardziej przypomina sposób działania ludzkiego mózgu, który wyciąga wnioski na podstawie spontanicznej obserwacji i intuicji. Wraz ze zwiększaniem

się rozmiaru zbiorów danych prezentowane wnioski są coraz bardziej precyzyjne. Poniżej przykłady wykorzystania:

- analiza koszyka zakupowego;
 - wykrywanie anomalii;
 - rozpoznawanie podobnych obiektów.
- Reinforcement Learning (ang. Uczenie Wzmocnione) - maszyna otrzymuje gotowy zestaw dozwolonych działań, reguł i stwierdzeń oraz wykorzystuje je w taki sposób, aby osiągnąć pożądaną efekt. Można to porównać do nauki gry np. w darta. Zasady określające, ile punktów musi zdobyć zawodnik oraz fakt zakończenia wartości podwójną pozostają niezmiennie. Natomiast najbardziej optymalna kombinacja punktów otrzymanych z maksymalnie trzech rzuconych lotek zależy od indywidualnej decyzji gracza. Przykłady zastosowań to:
 - nawigacja GPS (wybór trasy bazując na danych o natężeniu ruchu i pogodzie);
 - przemysł gamingowy (dopasowanie scenariuszy rozgrywki do działań gracza);
 - robotyka (dostosowanie natężenia pracy robotów do popytu).

Opisane techniki zilustrowano również na rysunku 9:



Rysunek 9: Podstawowe techniki Uczenia Maszynowego(Mamczur, 2019)

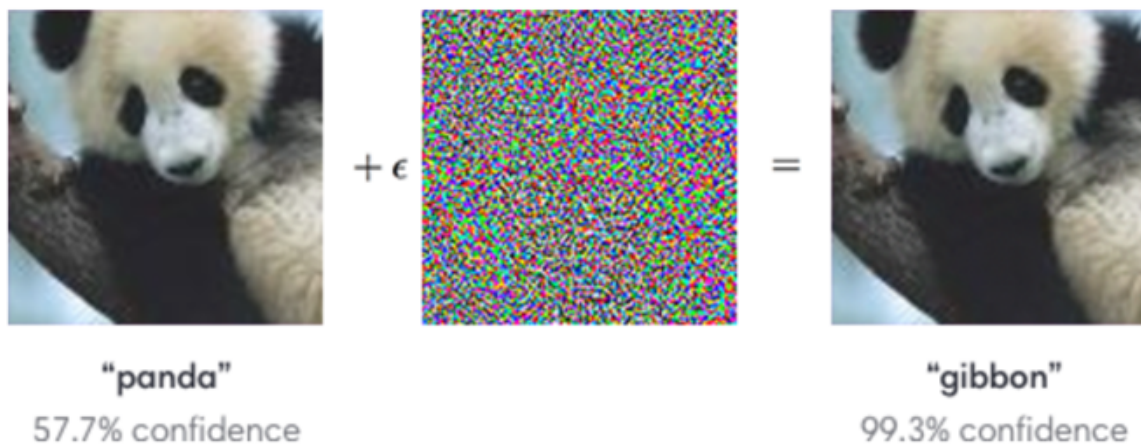
2.2 Charakterystyka Wrogiego Uczenia Maszynowego

Adversarial Machine Learning (AML), tłumaczone na język polski jako Kontradykcyjne Uczenie Maszynowe, czy też Wrogie Uczenie Maszynowe to dziedzina badań, która koncentruje się na opracowywaniu modeli ML odpornych na ataki kontradykcyjne, które są stosowane do oszukiwania lub manipulowania modelami uczenia maszynowego przez imputację złośliwych lub wprowadzających w błąd danych podczas faz uczenia lub wnioskowania. Ataki te mogą mieć poważne konsekwencje, od kradzieży poufnych informacji po nieprawidłowe działanie krytycznych systemów, takich jak samojezdne samochody lub urządzenia medyczne.

W celu dokładniejszego opisanie zjawiska AML, w kolejnych akapitach odwoływano się nie do ataków na konkretne systemy Uczenia Maszynowego, lecz w ujęciu ogólnym – jako ataki na Sztuczną Inteligencję (AI). W tej tematyce wkraczamy również w dziedzinę bezpieczeństwa w cyberprzestrzeni. W sektorze cyberbezpieczeństwa AML próbuje oszukać modele poprzez tworzenie unikalnych, wprowadzających w błąd danych wejściowych, aby zmylić model, powodując jego wadliwe działanie(zephyrnet.com, 2022).

Przeciwnicy (w języku angielskim nazywani są jako „attackers”, tłumaczone na język polski również jako „napastnicy”, czy też „adwersarze”) mogą wprowadzać dane, które mają za zadanie zmanipulować rezultaty wyjściowe, wykorzystując luki w modelu. Nie jesteśmy w stanie zidentyfikować tych danych wejściowych ludzkim okiem, jednak powoduje to nieprawidłowe działanie modelu. W systemach AI występują różne formy podatności, takie jak tekst, pliki audio, obrazy. Dużo łatwiej przeprowadzać ataki cyfrowe, takie jak manipulowanie tylko jednym pikselem w danych wejściowych, co może prowadzić do błędnej klasyfikacji(zephyrnet.com, 2022).

Przykład manipulacji obrazu przedstawiono na rysunku 10. W tym przypadku atakowano system rozpoznawania zwierząt, który został nauczony na bazie pewnej puli zdjęć. Przed atakiem, model rozpoznał, że na fotografii znajduje się panda, określając to z pewnością bliską 58%. Po dodaniu szumu, zmanipulowano system do tego stopnia, iż z niemal 100% przekonaniem sklasyfikował zdjęcie pandy jako gibbona(openai.com, 2017). Łatwo zauważyć, że jednym zaburzeniem napastnik zmienił status modelu z przydatnego na bezużyteczny.



Rysunek 10: Przykład ataku na system rozpoznawania obrazów (openai.com, 2017)

O ile zmanipulowanie systemu rozpoznawania obrazów zwierząt może wydawać się niegroźne i niedostatecznie ukazywać niebezpieczeństwo płynące z tego typu ataków, o tyle wpływ napastników np. na działanie samochodów autonomicznych wskazuje na tragiczne skutki jakie mogą zostać spowodowane. Jednym z mniej skomplikowanych algorytmów ML zastosowanych w tych środkach transportu jest system rozpoznawania znaków drogowych, jako że ich liczba jest skończona i względnie nieduża, a ich kształt, kolor i rozmiar jest ściśle znormalizowany. Dla przykładu rozważmy typową sytuację drogową.

Na rysunku 11 zamieszczono widok z przedniej kamery samochodu autonomicznego, tuż przed skrzyżowaniem. Auto bez problemu rozpoznaje znak STOP, a następnie wykonuje odpowiednie czynności, aby zatrzymać się przed skrzyżowaniem. Jednakże bardzo łatwo jest wprowadzić system w błąd, co może wydarzyć się na skutek zabrudzenia znaku,



Rysunek 11: Widok z kamery przedniej samochodu autonomicznego. Właściwie rozpoznany znak STOP (Surma, 2020)



Rysunek 12: Widok z kamery przedniej samochodu autonomicznego. Niewłaściwie rozpoznany znak STOP (Surma, 2020)

czy pomalowania go farbą, czego przykład przedstawiono na rysunku 12(Surma, 2020). Wskutek tego typu zaburzenia danych wejściowych, system oparty na Uczeniu Maszynowym może nie tylko nie rozpoznać tego znaku jako nakaz zatrzymania się, a wręcz może przypisać do otrzymanego obrazu zupełnie inny znak, mający w danym momencie krytyczne znaczenie dla bezpieczeństwa kierowcy. Na rysunku 13 przedstawiono przykładową interpretację przez model ML, gdzie zabrudzony znak STOP został przyjęty jako znak pierwszeństwa(Surma, 2020).

Zakładając, że aby dotrzeć do celu kierowca na danym skrzyżowaniu musi skrócić w lewo, samochód bez zatrzymywania się przejedzie przez to skrzyżowanie. Skutki takiej decyzji z wysokim prawdopodobieństwem będą tragiczne, co ukazuje jak duże znaczenie



Rysunek 13: Widok z kamery przedniej samochodu autonomicznego. Zakładając, że aby dotrzeć do celu kierowca na danym skrzyżowaniu musi skrócić w lewo, samochód bez zatrzymywania się przejedzie przez to skrzyżowanie. Skutki takiej decyzji z wysokim prawdopodobieństwem będą tragiczne, co ukazuje jak duże znaczenie

ma jakość danych dostarczanych do modelu i jak niewielkie zaburzenie może wpłynąć na jego niezawodność, a co za tym idzie, uzasadnienie dalszego wykorzystania zaprojektowanego narzędzia w biznesie(Surma, 2020).

Jednym z głównych czynników blokujących wykorzystanie Sztucznej Inteligencji w coraz to istotniejszych aspektach życia jest jej wrażliwość na wrogą ingerencję. Opisany powyżej przypadek obrazuje problemy z jakimi spotykają się współcześni modelarze systemów uczących się. Wyzwaniem najbliższych lat jest uczynienie tego typu narzędzi bezpiecznymi dla codziennego użytku. Aktualnie wciąż trwa wypracowywanie najodpowiedniejszych technik wzmacniania niezawodności ML, jak również definiowane są tzw. dobre praktyki, będące tymczasową odpowiedzią na niemoc badaczy w niektórych obszarach.

Badacze z instytutu naukowo-badawczego w Albuquerque jako cel obrali sobie stworzenie w pełni wiarygodnego i skutecznego sposobu na obronę przed wrogą ingerencją w modele Uczenia Maszynowego. Aby tego dokonać, przeprowadzili szereg badań nad wieloma zbudowanymi przez nich sieciami neuronowymi ukierunkowanymi na typowe dla tej tematyki rozpoznawanie obrazów. W pierwszym kroku wytrenowali modele próbując osiągnąć możliwie najwyższą skuteczność. Kolejnym etapem było obniżanie ich jakości poprzez "zatrutowanie" danych stosowanych do uczenia tychże modeli. Finalnie porównywano stopień degradacji skuteczności klasyfikatorów(Short, Pay, & Gandhi, 2019).

W celu podniesienia odporności modeli, stosuje się trening kontradyktoryjny, polegający na wprowadzaniu do modelu mylących danych, mających za zadanie wprowadzenie algorytmu w błąd, jednakże tego typu informacje, wprowadzone na etapie uczenia systemu, cechują się potencjałem do zmniejszania jego wrażliwości na ataki. Zatem można przyjąć, że poza technikami defensywnymi, równie szerokim zagadnieniem jest tworzenie strategii ataku. W literaturze znajdowane są odniesienia do kilku głównych metod kreacji wrogich danych. Niektóre z nich to(Short et al., 2019):

- Fast Gradient Sign Method - celem trenowania modelu jest minimalizacja tzw. funkcji błędu, natomiast poprzez wprowadzenie FGSM zwiększana jest wartość tejże funkcji;
- Basic Iterative Method - metoda będąca wprost rozwinięciem techniki FGSM, polegająca na wielokrotnym, aczkolwiek jasno wcześniej zdefiniowanym, jej powtórzeniu;

- Metoda Carliniego Wagnera (C&W) - technika generowania wrogich danych, skupiająca się na maksymalizacji podobieństwa między oryginalnymi informacjami wejściowymi, a zmanipulowanymi, zachowując przy tym efekt błędnej klasyfikacji przez model.

Naukowcy z Albuquerque wskazują wysoki potencjał zauważony przez nich w metodzie C&W. Mimo wielu prób użycia odpowiedniej taktyki defensywnej, badacze nie byli w stanie zastosować skutecznej obrony przed wrogimi danymi wygenerowanymi w ten sposób (Short et al., 2019), co podkreśla jak duże pole do dalszych badań istnieje w dziedzinie AML.

2.3 Typy ataków na algorytmy Uczenia Maszynowego

Na daną chwilę większość algorytmów proponowanych przez badaczy, naukowców i specjalistów z branży R&D skupia się głównie na wysokiej wydajności i niskiej liczbie błędnych klasyfikacji. Jednakże nawet kiedy wskazane cele zostają osiągnięte, modele te często nie powinny być implementowane w środowiskach produkcyjnych, zwłaszcza w domenach krytycznych, czy aspektach życia, które mogą mieć wpływ na życie znacznej części społeczeństwa, nie uwzględniając innych kryteriów i wymagań dotyczących sztucznej inteligencji. Są nimi: bezpieczeństwo algorytmów, ich interpretowalność i uczciwość. Co więcej, rezultaty osiągane są na danych, które są odpowiednio przygotowane w warunkach laboratoryjnych i możliwe jedynie w sytuacji gdy implementacja również zachodzi w takich warunkach (Pawlicki, 2020).

Zastosowanie Sztucznej Inteligencji na wielką skalę stało się rzeczywistością, za czym idzie świadomość, że bezpieczeństwo samych algorytmów Uczenia Maszynowego wymaga natychmiastowej uwagi. Adwersarze potrafią starannie dobrać próbki danych wejściowych, aby zmieniało to wyniki klasyfikacji lub regresji w oczekiwany przez nich sposób. Świadomość zagrożeń związanych z ich użyciem, a także ich podatność na AML jest jeszcze całkiem niewielka (Pawlicki, 2020), jednakże już powstały definicje określające poszczególne typy ataków.

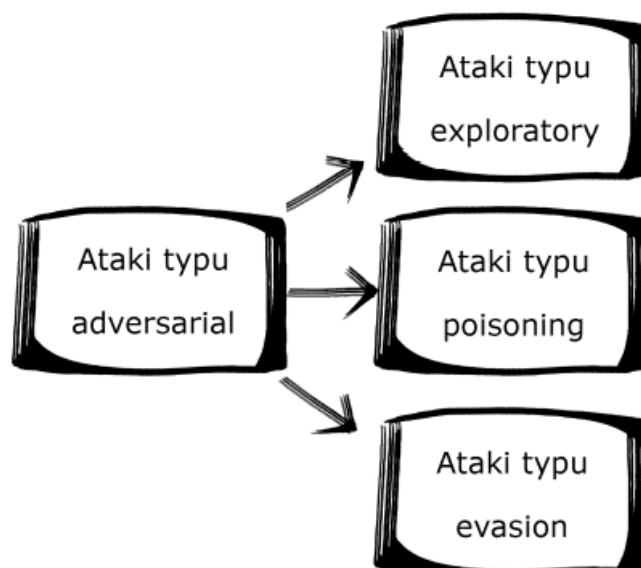
Jednym z najbardziej znanych podziałów jest klasyfikacja ze względu na dostęp do modelu (zephyrnet.com, 2022):

- Atak białoskrzynkowy - odnosi się do sytuacji, w której atakujący ma pełny dostęp

do modelu docelowego. Obejmuje to architekturę i parametry, które pozwalają im tworzyć próbki danych na modelu docelowym. Osoby atakujące będą miały ten dostęp tylko wtedy, gdy testują model jako programista. Mają oni detaliczną wiedzę na temat architektury sieci oraz znają tajniki modelu i tworzą strategię ataku;

- Atak czarnoskrzynkowy - odnosi się do sytuacji, w której atakujący nie ma dostępu do modelu docelowego i może jedynie zbadać dane wyjściowe.

Podział ataków na czarnoskrzynkowe i białoskrzynkowe oparty jest o umiejscowienie atakującego. Inna klasyfikacja bazuje na strategii ingerencji w model, a jej schemat przedstawiono na rysunku 14. Atak zatruwający (typu poisoning) koncentruje się na danych ze



Rysunek 14: Nowe ataki na Uczenie Maszynowe (Pawlicki, 2020)

zbioru uczącego. Tutaj atakujący zmienia istniejące lub wprowadza nieprawidłowo oznakowane dane. Wskutek takiego działania, model przeszkolony na „zatrutym” zbiorze będzie tworzył błędne predykcje na prawidłowo oznakowanych danych (towardsdatascience.com, 2021). W literaturze znaleźć można kilka artykułów na temat ataków tego typu. W jednym z nich, autorzy opisują użycie tzw. wrogiego szumu etykiet (ang. adversarial label noise). W tym artykule przedstawiona jest metoda wykorzystująca sposób działania algorytmu SVM (Support Vector Machines – ang. Metoda Wektorów Nośnych), którego działanie polega na mapowaniu danych na wielowymiarową przestrzeń właściwości w sposób umożliwiający kategoryzację punktów danych (ibm.com, 2021).

Ogólnym założeniem ataku jest wprowadzenie do zbioru treningowego próbki, która

znacząco zmieni wynik klasyfikacji, obniżając skuteczność modelu. Taką próbkę można stworzyć poprzez rozwiązanie problemu optymalizacyjnego, polegającego na wyszukiwaniu lokalnych maksimów powierzchni funkcji błędu, do czego wykorzystano algorytm gradient ascent. Atak wykorzystuje odwracanie etykiet konkretnych próbek w klasyfikacji binarnej zbioru uczącego, przy założeniu, że dane w zbiorze walidacyjnym nie są w żaden sposób zmieniane (Biggio, Nelson, & Laskov, 2012).

Ataki unikowe (typu evasion), w odróżnieniu od ataków zatruwających, nie skupiają się na danych używanych do uczenia modelu, lecz na odpowiedniej manipulacji danymi wejściowymi, dla których model wydaje prognozowany rezultat. Polegają one na modyfikowaniu danych, aby wydawały się uzasadnione, lecz by prowadziły do błędnej prognozy. Przykładem wykorzystania tego typu ataków są modele oceny wiarygodności kredytowej. Ubiegając się o kredyt, osoba atakująca może zamaskować swój prawdziwy kraj pochodzenia za pomocą usługi VPN, ukrywając w ten sposób np. iż jest obywatelem kraju uznawanego przez model jako bardziej ryzykowny, co mogłoby zmniejszyć jego szanse na pozytywną ocenę jego wniosku (towardsdatascience.com, 2021).

Innym kierunkiem wykorzystania ataków unikowych są modele służące do odfiltrowywania wiadomości e-mail będących spamem. Ich podejście może polegać na eksperymentowaniu z mailami, które model już wytrenował w zakresie sprawdzania i rozpoznawania jako spam. Jeśli model został wyszkolony do filtrowania wiadomości e-mail zawierających konkretne słowa, atakujący może tworzyć nowe e-maile zawierające powiązane z tym słowa, które przejdą przez algorytm, co spowoduje, że wiadomość e-mail, która w typowym procesie zostałaby sklasyfikowana jako spam, spamem nie jest, co w oczywisty sposób pogarsza skuteczność modelu (zephyrnet.com, 2022).

Atak poszukiwawczy (typu exploratory) może wystąpić po wytrenowaniu algorytmu, a jego zadaniem jest odkrywanie informacji o wewnętrznym działaniu modelu, w celu identyfikacji słabych punktów. W tym podejściu ingerencja jest ukierunkowana na poszukiwanie informacji o (Shi, Sagduyu, & Grushin, 2017):

- granicy decyzyjnej używanej przez algorytm (np. hiperpłaszczyzny maszyny wektorów nośnych (SVM) algorytm);
- ogólnym zestawie reguł, którymi kieruje się algorytm;

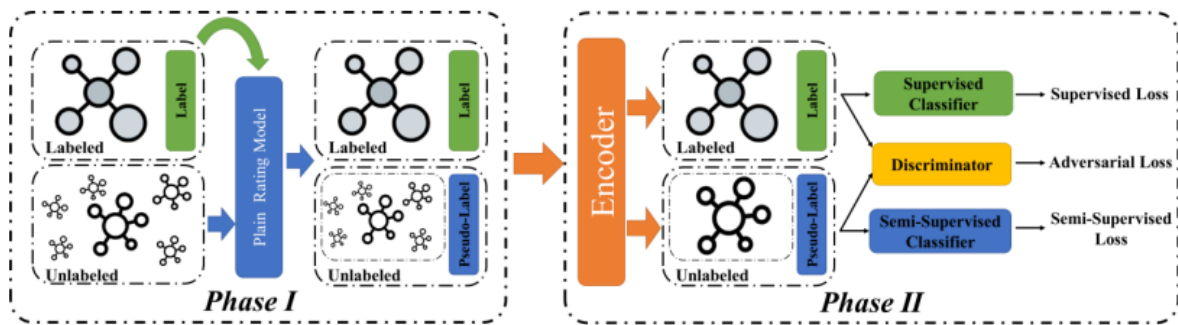
- zestawie logicznych lub probabilistycznych właściwości algorytmu;
- danych, które zostały wykorzystane (lub nie wykorzystane) do uczenia algorytmu.

W ostatnich latach powstało kilka prac, badających ataki na Głęboką Sieć Neuronową (DNN). W artykule z 2017 roku (Shi et al., 2017) naukowcy zbudowali tzw. funkcjonalny ekwiwalent klasyfikatorów modelu DNN, opierając się na algorytmach SVM oraz naiwnego klasyfikatora Bayes’a. Z kolei w badaniu z 2019 roku (Shi, Sagduyu, Davaslioglu, & Li, 2019) opisano jak przy pomocy Głębokiej Sieci Neuronowej można wydobyć klasyfikator wdrożony w warunkach produkcyjnych. Podsumowując - celem ataku typu exploratory jest budowa lokalnego klasyfikatora (Pawlicki, 2020).

Jak zatem zabezpieczyć się przed wrogimi ingerencjami w model? Jedną z możliwości stanowi tzw. trening kontradyktoryjny, będący jednym z podstawowych podejść do poprawy wydajności i bezpieczeństwa ML. Polega on na kontrolowanym atakowaniu modelu na wiele sposobów, znajdując w ten sposób „czułe punkty” zastosowanego rozwiązania. Na skutek takich badań, osoby budujące model mogą zweryfikować jego odporność na wrogie ataki, a następnie podjąć odpowiednie kroki celem poprawy jego bezpieczeństwa (zephyrnet.com, 2022).

Chińscy badacze postanowili wykorzystać zjawisko Adversarial Machine Learning do wyznaczania korporacyjnych rating’ów kredytowych. Dotychczas wykorzystanie standardowych metod predykcyjnych nie przynosiło satysfakcjonujących rezultatów, a ma na to wpływ specyfika rynku. Zatrudnienie i opłacenie zespołu specjalistów, których zadaniem będzie oszacowanie stopnia wiarygodności firmy nie jest małym wydatkiem i mogą sobie na nie pozwolić jedynie duże korporacje. Implikuje to względnie niedużą próbę obserwacji i możliwość ich uzasadnionego wykorzystania jedynie dla podobnych instytucji, jak te posiadające swój rating. Większość średnich i praktycznie wszystkie małe przedsiębiorstwa nie są w stanie otrzymać własnego rate’u, a co za tym idzie, potencjalny model nie ma na czym nauczyć się, a potem przewidywać wiarygodność tych firm (Feng & Xue, 2021).

Naukowcy zaproponowali kilkuetapowy proces uczenia modelu, zobrazowany na rysunku 15, rozpoczynający się od podzielenia danych na oznaczone (ang. labeled data) - firmy ze znanym rating’iem, a także nieoznaczone (ang. unlabeled data) - firmy bez rating’u. Na danych oznaczonych nauczono model predykcyjny gradientowo wzmocnionego drzewa decyzyjnego, który następnie wykorzystano do wyznaczenia najbardziej prawdopo-



Rysunek 15: Proces uczenia według ASSL4CCR (Feng B., 2021)

dobnych rating'ów dla obserwacji nieoznaczonych. W ten sposób otrzymano pełen zestaw niewybrakowanych danych, jednakże kluczowym elementem tego etapu było zakwalifikowanie ocen wyliczonych przez model jako 'pseudooceny' - rate'y z ograniczonym do nich zaufaniem(Feng & Xue, 2021).

Proces ten można porównać do nauki ucznia w szkole. W trakcie lekcji, otrzymuje on zestaw zadań i poprawnych do nich odpowiedzi, które dostarcza nauczyciel, a zatem można w pełni zaufać, że są one zgodne z rzeczywistością. Jednakże podczas sprawdzianu, gdy pojawią się zadania nieco inne niż podczas zajęć, podejrzenie odpowiedzi u kolegi daje informację o tym jak do danego pytania ustosunkowała się inna osoba, ale nie musi to oznaczać, że rozwiązanie "ściągnięte" od sąsiada jest zgodne z rzeczywistością, więc trzeba je traktować z ograniczonym zaufaniem(Feng & Xue, 2021).

Model	Recall	Accuracy	F1-score
LR	0.76250	0.80970	0.81946
SVM	0.83750	0.89247	0.88961
MLP	0.91406	0.93568	0.93245
Xgboost	0.92343	0.94225	0.94133
CCR-CNN	0.92812	0.95253	0.94518
CCR-GNN	0.93437	0.95012	0.95177
ASSL4CCR	0.95321	0.96115	0.96252

Tabela 3: Klasyczna karta oceny punktowej(Feng B., 2021)

W drugim etapie na przygotowanych danych zastosowano podejście z kategorii AML, polegające na ograniczeniu zaufania do danych wypredykowanych z obserwacji nieoznaczonych, dzięki czemu mogą one brać udział w analizie. Na podstawie zbiorów danych oznaczonych oraz nieoznaczonych obliczono klasyfikator nadzorowany i klasyfikator częściowo-nadzorowany, a dodatkowo wyznaczono kontradyktoryjną stratę między zbiorami co po-

zwoliło na zastosowanie obydwu zbiorów jako całości i wyznaczenie modelu opisanego jako ASSL4CCR - Kontrykcyjne Uczenie Częściowo Nadzorowane dla Korporacyjnego Rating'u Kredytowego(Feng & Xue, 2021).

Chińscy naukowcy, jako zwieńczenie projektu, zrealizowali porównanie podstawowych parametrów pozwalających na obiektywną ocenę jakości poszczególnych modeli tj. recall, dokładność oraz F1-score. Z analizy wyników przedstawionych w tabeli 3 wynika, że model ASSL4CCR osiąga najlepsze rezultaty, co dowodzi przydatności świadomego wykorzystania AML w praktyce(Feng & Xue, 2021).

2.4 Przykłady ataków na algorytmy Uczenia Maszynowego

Adversarial Machine Learning jest względnie nową poddziedziną Analizy Danych i na tę chwilę nie łatwo jest znaleźć wzięte z życia przykłady ingerencji w rzeczywiste, używane produkcyjnie modele ML. Dzięki wysokiej świadomości analityków nie trzeba uczyć się na nieświadomie popełnionych błędach następnie wykorzystanych przez adwersarzy, a problem został zidentyfikowany zanim pojawiły się jego potencjalnie fatalne skutki. W związku z powyższym, w ostatnich latach, naukowcy prowadzą intensywne badania, samodzielnie generując różne scenariusze ingerencji w model, jednakże wciąż pozostaje w tej dziedzinie wiele przestrzeni dla nowych odkryć.

W ostatnich latach prowadzone jest coraz więcej badań na modelach wykorzystujących Uczenie Maszynowe pod kątem poprawienia ich odporności na zamierzone, bądź też przypadkowe zaburzenia i ataki. Ze względu na względnie wysoką podatność oraz szeroką gamę praktycznych zastosowań najbardziej interesującym, a zarazem najczęściej podejmowanym obiektem rozważań są sieci neuronowe i głębokie sieci neuronowe(Papernot, McDaniel, Goodfellow, Jha, & Celik, 2017).

Większość ciekawych wniosków dotyczy prac w tematyce rozpoznawania obrazów, a konkretnie uodparniania algorytmów na potencjalne ataki. Na łamach cytowanej pracy, naukowcy z Uniwersytetu stanowego Pensylwanii oraz Uniwersytetu w Wisconsin, przy współpracy z reprezentantem firmy OpenAI, która najbardziej znana jest ze stworzenia ChatuGPT(openai.com, 2023), zajęli się praktycznymi atakami na ogólnodostępne modele głębokich sieci neuronowych przy pomocy tzw. ataków czarnokrzynekowych (ten jak i również pozostałe typy ataków opisano w podrozdziale 2.3). Badacze zrealizowali serię



Rysunek 16: Przykłady obrazów wykorzystanych w ataku na model głębokiej sieci neuronowej firmy MetaMind(Papernot et al., 2017)

testów na oprogramowaniu firmy MetaMind, obserwując odpowiedzi modelu na wprowadzane przez nich obrazy. Następnie wygenerowali tzw. Wrogie Przykłady (ang. Adversarial Examples), nazywane dalej Wrogimi Próbkami, które cechują się niezauważalnymi lub trudno zauważalnymi dla człowieka różnicami, w przypadku plików graficznych na poziomie manipulacji pojedynczych pikseli. Na rysunku 16 wskazano przykładowe obrazy wykorzystane w eksperymencie. W górnym wierszu przedstawiono poprawnie sklasyfikowane przykłady, natomiast niżej zaprezentowano przygotowane i błędnie ocenione przez model Wrogie Próbki(Papernot et al., 2017).

Wyniki badań wskazywały, iż naukowcy są w stanie zaburzyć klasyfikacje generowane przez głęboką sieć neuronową w 84,24% przypadków. Jeszcze więcej do myślenia dają kolejne kroki opisane przez naukowców. Stosując to samo podejście, przetestowali skuteczność klasyfikatora zaimplementowane w dostępnych online rozwiązaniach firm Amazon oraz Google. Zarejestrowano błędne klasyfikacje modeli w odpowiednio 96,19% oraz 88,94% przypadków(Papernot et al., 2017).

Obecnie lwia część społeczeństwa posiada skrzynkę mailową, a coraz częściej jedna osoba posiada kilka takich komunikatorów. Łatwość tworzenia nowych kont i docierania do innych użytkowników na masową skalę, wciąż stanowi bardzo popularny sposób nie tylko do porozumiewania się w ważnych sprawach tj. komunikacja służbowa czy też potwierdzenia różnych transakcji internetowych, lecz także daje możliwość przesyłania różnych ofert i reklam. Wśród takich wiadomości łatwo jest umieszczać niebezpieczne linki, kie-

rujące nieświadomych odbiorców do domen stanowiących zagrożenie dla ich prywatności. Według raportu Message Labs Intelligence pośród całej globalnej komunikacji mailowej, wiadomości typu spam stanowią aż 88%(Kuchipudi, Nannapaneni, & Liao, 2020). W celu uniknięcia podejrzanych wiadomości oraz zalewaniu skrzynki przez nachalną propagandę reklamową stosuje się tzw. filtry antyspamowe oparte na algorytmach uczenia maszynowego.

Algorytmy uczone są wykrywania "wrogich" słów w zawartości e-maili. W zależności od modelu różne słowa w zróżnicowanym stopniu obniżają lub podnoszą poziom zaufania modelu do danej wiadomości, a po przeprowadzeniu oceny podejmowana jest decyzja o umieszczeniu jej w skrzynce odbiorczej lub folderze spam. Na polu filtracji niechcianych wiadomości toczona jest ciągła walka między deweloperami, a adversarzami mającymi na celu oszukać model, aby dokonał błędnej klasyfikacji(Cheng, Xu, Li, & Ding, 2022).

Jednym ze sposobów wykorzystywanych przez napastników jest zaszycie w wiadomości dostatecznie wielu silnie zaufanych słów, które przeważą ocenę klasyfikatora z negatywnej na pozytywną. Jednakże umieszczenie wielu słów z bardzo wąskiej i zróżnicowanej grupy może skutecznie wypaczyć przekaz wiadomości do tego stopnia, że będzie ona łatwo wykrywalna nawet dla ludzkiego oka, a potencjalna ofiara nie da się nabrać na atak. Pomysłów adversarzy na uniknięcie takiej sytuacji jest wiele i odwołują się one głównie ich kreatywności, jak np. wpisanie kilkudziesięciu zaufanych słów w zawartość maila, stosując jednocześnie dla nich kolor czcionki identyczny z barwą tła, co skutecznie ukrywa podejrzaną zawartość zarówno przed użytkownikiem, jak i algorytmem, czy też umieszczenie literówek w słowach znacznie obniżających prognozowaną przez algorytm ocenę takiej zawartości, dzięki czemu uzyskują one wagę neutralną i nie są wychwytywane przez filtry jako niebezpieczne(Cheng et al., 2022).

Zespół naukowców z Uniwersytetu Michigan postanowił przeprowadzić badania mające na celu sprawdzić skuteczność ataków typu AML na algorytmy anty-spamowe. Wykorzystali do tego celu trzy względnie proste techniki(Kuchipudi et al., 2020):

- synonym replacement (ang. podmiana słowa przy pomocy synonimu);
- ham word injection (ang. wprowadzenie zaufanych słów);
- spam word spacing (ang. wprowadzenie przerw/spacji pomiędzy literami w słowach)

związanych ze spamem).

Modelem, którego skuteczność do obrony badano, był algorytm naiwnego klasyfikatora Bayesa. Badacze argumentują wybór właśnie tej techniki implementacji filtra antyspamowego jego popularnością w tego typu zastosowaniach, co empirycznie wykazało jego wysoką skuteczność w wychwytywaniu wrogich, bądź też niechcianych wiadomości (Kuchipudi et al., 2020).

Naukowcy na łamach pracy wskazują również potencjalne trudności, jakie mogą wystąpić przy próbie tego typu ataków. Główną przeszkodą może być "czarnoskrzynkowość", czyli ograniczony lub całkowicie uniemożliwiony dostęp do informacji na temat wybranego i nauczonego modelu filtracji antyspamowej, jak również nieznaną doświadczeń danych wejściowych użytych do jego budowy. Jednakże warto zauważyć, że badania wykazują, iż do przeprowadzenia udanego ataku na algorytmy uczenia maszynowego filtrów spam, często wystarczy znajomość zaledwie 1% danych treningowych, co znacznie ułatwia zadanie adwersarzy (Kuchipudi et al., 2020).

Modified Message	Cosine Similarity	Prediction
Ringtone Club: acquire the UK single graph on your Mobile_River each hebdomad and take any top_side caliber ringtone! This content is free_people of charge.	0.583	spam
Ringtone Club: become the UK bingle graph on your nomadic each workweek and select any upper_side caliber ringtone! This subject_matter is liberate of charge.	0.583	spam
Ringtone Club: go the UK one graph on your peregrine each calendar_week and pick_out any upside character ringtone! This substance is release of charge.	0.583	ham

Tabela 4: Wrogie próbki zastosowane na atakowanym modelu filtra antyspamowego (Kuchipudi et al., 2020)

W pierwszym podejściu, zastosowano synonimy słów, wychwytywanych przez filtr jako niebezpieczne, nie zmieniając przy tym sensu samej wiadomości. Wykorzystano do tego celu technikę NLP - Natural Language Processing(ang. przetwarzanie języka natural-

nego) - będącą poddziedziną Sztucznej Inteligencji i odpowiedzialną za rozumienie języka ludzkiego przez maszyny i roboty(Castagno, 2020). Na łamach pracy, jako przykład przetworzenia wiadomości niebezpiecznej zmanipulowanej w sposób pozwalający przejść przez filtrację podano zdanie: "Ringtone Club: Get the UK singles chart on your mobile each week and choose any top quality ringtone! This message is free of charge." Różnie zmodyfikowane wiadomości wraz z rezultatem ich klasyfikacji przez model antyspamowy przedstawiono w tabeli 4(Kuchipudi et al., 2020). Łatwo zauważyć, że wiadomość, która została przepuszczona przez klasyfikator jako zaufana, zapewne nie zostałaby potraktowana poważnie przez użytkownika, co wskazuje na wdrożenie niezbędnych poprawek w zestawach synonimów, wykluczając te mające wpływ na kontekst informacji.

Drugie podejście opiera się na manipulacji częstotliwością pojawiania się słów zaufanych. Wśród publicznie dostępnych zbiorów można bez trudu znaleźć zestawy słów znanych jako słowa związane ze zjawiskiem spamu i z wysokim prawdopodobieństwem generujące uruchomienie filtra. W związku z powyższym, jako zaufane należy traktować wyrazy, które nie znajdują się w tym zbiorze, dzięki czemu wprowadzenie ich do zawartości wiadomości podnosi prawdopodobieństwa przejścia przez filtrację(Kuchipudi et al., 2020).

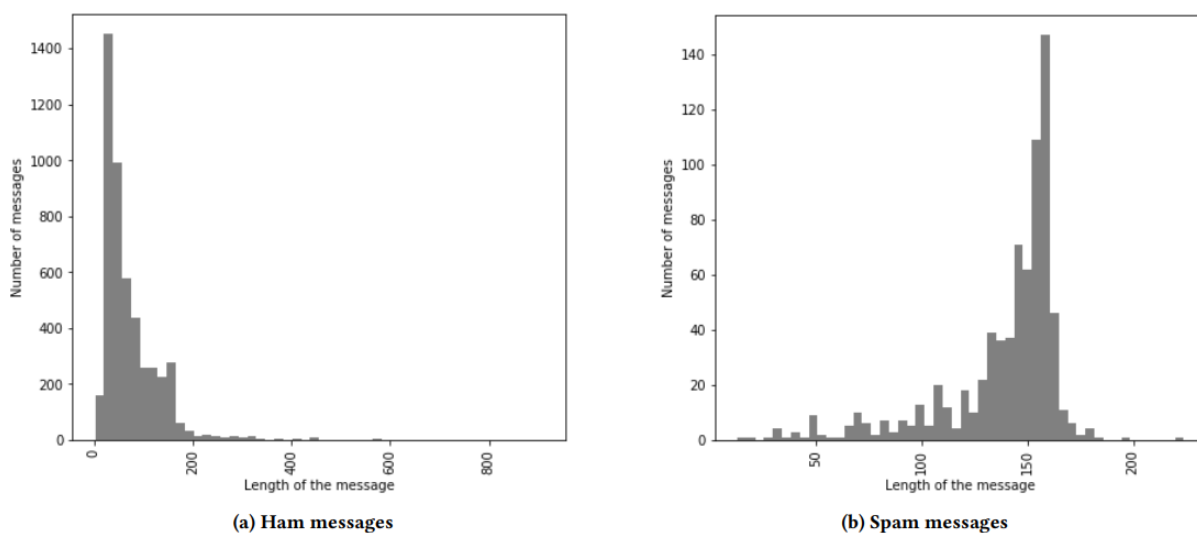
Przykładowa wiadomość klasyfikowana początkowo jako spam: "Congratulations ur awarded 500 of CD vouchers or 125gift guaranteed and Free entry 2 100 wkly draw txt MUSIC to 87066 TnCs www.Ldew.com1win150ppmx3age16", po wprowadzeniu kilkukrotnie słów zaufanych tj. good, love, deal, jest w stanie przejść przez filtr jako zaufana: "Congratulations good ur awarded good 500 of CD vouchers or 125 good gift guaranteed love and Free entry 2 good 100 wkly draw txt MUSIC to 87066 TnCs www.Ldew.com1win150ppmx3age16 good good good good good deal". Ponownie można zastanawiać się nad sensem samej wiadomości, jednakże nie było to obiektem rozważań cytowanej pracy (Kuchipudi et al., 2020).

Podczas badań tego podejścia zauważono pewną właściwość filtrów antyspamowych. Z testów można wnioskować, że model jest wyczulony na używanie skrótów w mailach tj. użycie "U" zamiast "you" (ang. Ty) czy też "R" w miejsce "are" (ang. jesteś/jest/jesteśmy/jesteście/są). Wynika z tego, że stosowanie bardziej zadbanego i oficjalnego języka tekstu daje większą szansę na ominięcie filtracji i skuteczne umieszczenie potencjalnie wrogiej

wiadomości w atakowanej skrzynce odbiorczej(Kuchipudi et al., 2020).

Trzecia technika polegała na wprowadzeniu odstępów między znakami w słowach prawdopodobnie klasyfikowanych jako niebezpieczne. Wysoki współczynnik podobieństwa, przy pomocy którego badano podobieństwo znaczeniowe wiadomości oryginalnej i zmanipulowanej, wskazuje iż w tym przypadku treść jest najmniej zaburzona względem odniesienia i możliwie najlepiej oddaje sens zamierzony przez autora. W przytoczonym przez badaczy przykładzie edytowanie słów "sexy"i "flirt"do form "s e x y"oraz "f l i r t"skutkuje oszukaniem modelu i uznaniem wiadomości niebezpiecznej za zaufaną (Kuchipudi et al., 2020).

W podsumowaniu pracy można znaleźć konkluzję, iż przy wykorzystaniu trzech opisanych powyżej sposobów potrafiąco w około 60 % przypadków oszukać filtr antyspamowy. W trakcie badań wykorzystano zbiór danych z witryny Kaggle, zawierający 5572 wiadomości, z czego 747 zostały oznaczone jako spam. Wielkość zestawu jest wystarczająca do budowy takiego mechanizmu, jednakże zdecydowanie pewniejszy model można byłoby uzyskać przy zebraniu nieco większej próbki do jego nauki, co przyczyniłoby się do zebrania bardziej wiarygodnych wniosków(Kuchipudi et al., 2020).



Rysunek 17: Dwa histogramy obrazujące rozkład liczby znaków wiadomości w zależności od liczby znaków dla treści spam oraz tekstów zaufanych(Kuchipudi et al., 2020)

Ciekawe spostrzeżenie, mogące mieć duży wpływ na skuteczność wykrywania niebezpiecznych maili, płynie z analizy długości(liczby znaków) w wiadomościach. Zauważono, że treści zaufane zawierają zwykle poniżej 150 znaków, gdzie w przypadku spamu obser-

wuje się koncentrację wokół tejże liczby. Zobrazowano to na rysunku 17, jednakże należy zwrócić uwagę na oś poziomą, która w skutek różnej skali dla obydwu wykresów może nieco zakłamywać skalę zróżnicowania obserwowanych wartości(Kuchipudi et al., 2020).

Na łamach pierwszych dwóch rozdziałów opisano tematykę Credit Scoringu, jak również Wrogiego Uczenia Maszynowego. Rozdziały 3 i 4 stanowią będą praktyczne połączenie tych dwóch zagadnień.

3 Budowa modelu drzewa decyzyjnego do celu Credit Scoring

3.1 Opis zbioru danych oraz użytych narzędzi i technologii

3.1.1 Tytuł podpodrozdziału

3.1.2 Tytuł podpodrozdziału

3.1.3 Tytuł podpodrozdziału

3.2 Budowa modelu

3.2.1 Tytuł podpodrozdziału

3.2.2 Tytuł podpodrozdziału

3.2.3 Tytuł podpodrozdziału

3.3 Analiza wyników

3.3.1 Tytuł podpodrozdziału

3.3.2 Tytuł podpodrozdziału

3.3.3 Tytuł podpodrozdziału

Podsumowanie rozdziału

4 Atak na opracowany model

4.1 Strategia badania odporności modelu

4.1.1 Tytuł podpodrozdziału

4.1.2 Tytuł podpodrozdziału

4.1.3 Tytuł podpodrozdziału

4.2 Implementacja wybranych technik ataku

4.2.1 Tytuł podrozdziału

4.2.2 Tytuł podrozdziału

4.2.3 Tytuł podrozdziału

4.3 Analiza wyników i weryfikacja hipotez badawczych

4.3.1 Tytuł podpodrozdziału

4.3.2 Tytuł podpodrozdziału

4.3.3 Tytuł podpodrozdziału

Podsumowanie rozdziału

Wnioski

Literatura

1. Bajek, R. (2011). Wykorzystanie metod eksploracji danych do budowy modeli scoringowych. Politechnika Śląska, Instytut Informatyki.
2. bankier.pl. (2007). Ryzykowny sektor. <https://www.bankier.pl/wiadomosc/Ryzykowny-sektor-1662142.html>, 03.12.2007.
3. bankier.pl. (2012). Tajemnicza liczba, czyli credit scoring. <https://www.bankier.pl/wiadomosc/Tajemnicza-liczba-czyli-credit-scoring-2512458.html>, 02.04.2012.
4. Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines.
5. bik.pl. (2022). Jak poprawić swoją zdolność kredytową? <https://www.bik.pl/poradnik-bik/jak-poprawic-swoja-zdolnosc-kredytowa>, 5.10.2022.
6. britannica.com. (2019). Dendral. <https://www.britannica.com/technology/DENDRAL>, 19.09.2019.
7. Bujak, L. (2008). Drzewa decyzyjne. <http://www.is.umk.pl/duch/Wyklady/CIS/Prace%20zalicz/08-Bujak.pdf>, 2008.
8. Caire, D., Barton, S., de Zubiria, A., Alexiev, Z., & Dyer, J. (2006). A handbook for developing credit scoring systems in a microfinance context. Washington, Development Alternatives, Inc.
9. calcxml.com. (2023). Financial calculators. <https://www.calcxml.com/do/credit-score-calculator-new?skn=results>, 2023.
10. Castagno, P. (2020). How to identify spam using natural language processing (nlp)? <https://towardsdatascience.com/how-to-identify-spam-using-natural-language-processing-nlp-af91f4170113>, 2020.
11. Cheng, Q., Xu, A., Li, X., & Ding, L. (2022). Adversarial email generation against spam detection models through feature perturbation. Information Security Institute, Johns Hopkins University, Baltimore, MD; Department of Computer Science, American University, Washington, D.C.
12. crif.pl. (2018). Rola „machine learning” w procesach kredytowych. <https://www.crif.pl/wiadomo%C5%9Bci/dla-prasy/2020/sierpie%C5%84/rola-machine-learning-w-procesach->

kredytowych/, 2018.

13. Deryło, L. (2021). Regresja logistyczna - co to jest? <https://www.lukaszderlylo.pl/blog/regresja-logistyczna.html>, 31.01.2021.
14. direct.money.pl. (2022). Co to jest scoring kredytowy? jak banki ustalają credit scoring i jakich używają systemów? <https://direct.money.pl/artykuly/porady/czym-jest-credit-scoring>, 18.01.2022.
15. ecomparemo.com. (2020). A brief history of credit scoring in the world. <https://www.ecomparemo.com/info/a-brief-history-of-credit-scoring-in-the-world>, 23.11.2020.
16. elektronikab2b.pl. (2021). Czym jest uczenie maszynowe i jak można je wykorzystać? <https://elektronikab2b.pl/biznes/53039-czym-jest-uczenie-maszynowe-i-jak-mozna-je-wykorzystac>, 11.12.2020.
17. experian.com. (2021). What is a good credit score? <https://www.experian.com/blogs/ask-experian/credit-education/score-basics/what-is-a-good-credit-score/>, 11.02.2021.
18. Fawcett, T. (2005). An introduction to roc analysis. Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306, USA.
19. Feng, B., & Xue, W. (2021). Adversarial semi-supervised learning for corporate credit ratings. Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing China, School of Artificial Intelligence, University of Chinese Academic of Science.
20. fico.com. (2023). Corporate information. <https://fico.gcs-web.com/corporate-information/>, 2023.
21. fotc.com. (2022). Machine learning — czym jest uczenie maszynowe? <https://fotc.com/pl/blog/machine-learning/>, 16.11.2022.
22. Gajowniczek, K., Ząbkowski, T., & Szupiluk, R. (2014). Estimating the roc curve and its significance for classification models' assessment. Warszawa, Department of Informatics, Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences - SGGW; Szkoła Główna Handlowa.
23. gov.pl. (2021). Co to jest uczenie maszynowe – inteligentna analiza da-

- nych? <https://www.gov.pl/web/popcwsparcie/co-to-jest-uczenie-maszynowe-inteligentna-analiza-danych>, 15.06.2021.
24. Group, T. W. B. (2021). Credit scoring approaches guidelines. Washington, The World Bank Group.
 25. habza.com.pl. (2022). Zdolność kredytowa a stopy procentowe. <https://habza.com.pl/zdolnosc-kredytowa-a-stop-y-procentowe/>, 15.06.2022.
 26. ibm.com. (2021). Sposób działania algorytmu svm. <https://www.ibm.com/docs/pl/spss-modeler/saas?topic=models-how-svm-works>, 17.08.2021.
 27. Karolak, Z. (2014). Dynamiczne ujęcie ryzyka kredytowego z ujęciem analizy przeżycia. https://www.sas.com/content/dam/SAS/pl_pl/image/events/mdb/prezentacje/zuzanna-karolak-dynamiczne-ujecie-ryzyka.pdf, 18.11.2014.
 28. Kuchipudi, B., Nannapaneni, R. T., & Liao, Q. (2020). Adversarial machine learning for spam filters. Department of Computer Science Central Michigan University Mt. Pleasant, Michigan, USA.
 29. Mamczur, M. (2019). Czym jest uczenie maszynowe? i jakie są rodzaje? <https://mirosławmamczur.pl/czym-jest-uczenie-maszynowe-i-jakie-sa-rodzaje/>, 30.11.2019.
 30. Matuszyk, A. (2009). Dotychczasowe oraz nowe trendy w metodzie "credit scoring".
 31. mfiles.pl. (2020). Credit scoring. https://mfiles.pl/pl/index.php/Credit_scoring, 19.05.2020.
 32. naukowiec.org. (2014). Regresja logistyczna - opis. https://www.naukowiec.org/wiedza/statystyka/regresja-logistyczna_466.html, 14.04.2014.
 33. newsblog.pl. (2022). Wyjaśnienie regresji a klasyfikacja w uczeniu maszynowym. https://newsblog.pl/wyjasnienie-regresji-a-klasyfikacja-w-uczeniu-maszynowym/Regresja_logistyczna, 1.12.2022.
 34. openai.com. (2017). Attacking machine learning with adversarial examples. <https://openai.com/research/attacking-machine-learning-with-adversarial-examples>, 24.02.2017.
 35. openai.com. (2023). About openai. <https://openai.com/about>, 2023.

36. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., & Celik, B. (2017). Practical black-box attacks against machine learning. ACM Asia Conference on Computer and Communications Security, Abu Dhabi, UAE.
37. Pawlicki, M. (2020). Zastosowanie metod uczenia maszynowego do wykrywania ataków sieciowych. Bydgoszcz, Uniwersytet Technologiczno-Przyrodniczy im. Jana i Jędrzeja Śniadeckich.
38. pl.economy.pedia.com. (2021). Punktacja kredytowa. <https://pl.economy.pedia.com/11030209-credit-scoring>, 2021.
39. Poon, M. (2007). Scorecards and devices for consumer credits: The case of fair, isaac and company incorporated. The Sociological Review, Issue Supplement S2, 55, s. 284–306.
40. prawnicydotblog.wordpress.com. (2019). Credit scoring. <https://prawnicydotblog.wordpress.com/2019/04/08/credit-scoring/>, 08.04.2019.
41. Przanowski, K. (2014). Credit scoring w erze big-data. Warszawa, Szkoła Główna Handlowa.
42. Przanowski, K. (2015). Credit scoring : Studia przypadków procesów biznesowych. Warszawa, Szkoła Główna Handlowa.
43. Przanowski, K. (2023). Credit scoring - automatyzacja procesu biznesowego - prezentacja do przedmiotu. Warszawa, Szkoła Główna Handlowa.
44. researchgate.net. (2017). 5 v's of big data. https://www.researchgate.net/figure/The-5V-of-Big-Data-Characteristics_fig1_321050765, październik 2017.
45. sas.com. (2018). Cztery typy uczenia maszynowego. https://www.sas.com/pl_pl/news/informacje-prasowe-pl/2018/cztery-typy-uczenia-maszynowego.html, 22.08.2018.
46. scoringexpert.pl. (2018). 8 ważnych informacji potrzebnych do zrozumienia nowej oceny punktowej, którą bik sprzedaje konsumentom. <http://scoringexpert.pl/2018/02/21/ocena-punktowa-bik-skala-do-100/>, 21.02.2018.
47. Shi, Y., Sagduyu, Y., Davaslioglu, K., & Li, J. (2019). Generative adversarial networks for black-box api attacks with limited training data.

48. Shi, Y., Sagduyu, Y., & Grushin, A. (2017). How to steal a machine learning classifier with deep learning. Rockville, MD 20855, USA, Intelligent Automation, Inc.,.
49. Short, A., Pay, T. L., & Gandhi, A. (2019). Defending against adversarial examples. y Sandia National Laboratories, operated for the United States Department of Energy by National Technology Engineering Solutions of Sandia, LLC.
50. Siddiqi, N. (2016). Credit risk scorecards developing and implementing intelligent credit scoring. New Jersey, John Wiley Sons, Inc.
51. Statsoft. (2010). Metody skoringowe w biznesie i nauce. https://media.statsoft.pl/_old_dnn/downloads/modele_skoringowe_w_biznesie.pdf, 2010.
52. StatSoft. (2010). Zastosowanie metod scoringowych w działalności bankowej. https://media.statsoft.pl/_old_dnn/downloads/zast_met_skoringowych_w_dz_bankowej.pdf, 2010.
53. statystyka.az.pl. (2021). Regresja logistyczna. <https://www.statystyka.az.pl/regresja-logistyczna.php>, 17.08.2021.
54. Surma, J. (2020). Prezentacja pt. hakowanie sztucznej inteligencji. Warszawa, Szkoła Główna Handlowa.
55. techtarget.com. (2021). 5 v's of big data. <https://www.techtarget.com/searchdatamanagement/definition/5-Vs-of-big-data>, marzec 2021.
56. Thomas, L., Edelman, D., & Crook, J. (2002). Credit scoring and its applications. Philadelphia, Society for Industrial and Applied Mathematics.
57. Thonabauer, G., & Nosslinger, B. (2004). Guidelines on credit risk management. credit approval process and credit risk management. Oesterreichische Nationalbank and Austrian Financial Market Authority.
58. totalmoney.pl. (2020). Ocena punktowa bik-u – jaka wartość scoringu bik-u jest dobra i daje szansę na kredyt? <https://www.totalmoney.pl/artykuly/179147,kredyty-gotowkowe,ocena-punktowa-bik-u—jaka-wartosc-scoringu-bik-u-jest-dobra-i-daje-szanse-na-kredyt,1,1,03.10.2020>.
59. towardsdatascience.com. (2021). What is adversarial machine lear-

- ning? <https://towardsdatascience.com/what-is-adversarial-machine-learning-dbe7110433d6>, 12.07.2021.
60. Weston, L. (2012). Your credit score. New Jersey, Pearson Education, Inc.
 61. Wikipedia. (2022). Fico. <https://en.wikipedia.org/wiki/FICO>, 24.12.2022.
 62. Wyśiński, P. (2013). Zastosowanie scoringu kredytowego w bankowości. Gdańsk, Uniwersytet Gdański.
 63. zephyrnet.com. (2022). Co to jest kontradycyjne uczenie maszynowe? <https://zephyrnet.com/pl/co-to-jest-kontradycyjne-uczenie-maszynowe/>, 3.03.2022.

Spis rysunków

1	Przykład cechy wykorzystanej w Credit Scoring(bankier.pl, 2012)	8
2	Diagram oceny wiarygodności kredytowej według firmy FICO(forbes.com, 2021)	10
3	Fragment kalkulacji ze strony calxml.com (calxml.com, 2023)	12
4	Wizualizacja oceny punktowej w BIK (źródło opracowanie własne)	12
5	Krzywa Profit zależna od mocy predykcyjnej(Przanowski, 2023)	17
6	Krzywa ROC i jej możliwe warianty(Gajowniczek et al., 2014)	18
7	Schemat 5V Big Data(researchgate.net, 2017)	20
8	Drzewo decyzyjne zastosowane do kategoryzacji potencjalnych kredytobiorców pod kątem ryzyka kredytowego (Bujak, 2008)	24
9	Podstawowe techniki Uczenia Maszynowego(Mamczur, 2019)	29
10	Przykład ataku na system rozpoznawania obrazów (openai.com, 2017) . . .	31
11	Widok z kamery przedniej samochodu autonomicznego. Właściwie rozpoznany znak STOP (Surma, 2020)	31
12	Widok z kamery przedniej samochodu autonomicznego. Niewłaściwie rozpoznany znak STOP (Surma, 2020)	32
13	Widok z kamery przedniej samochodu autonomicznego. Znak STOP błędnie rozpoznany jako znak bezwzględnego pierwszeństwa przy skręcie w lewo (Surma, 2020)	32
14	Nowe ataki na Uczenie Maszynowe (Pawlicki, 2020)	35
15	Proces uczenia według ASSL4CCR (Feng B., 2021)	38
16	Przykłady obrazów wykorzystanych w ataku na model głębokiej sieci neuronowej firmy MetaMind(Papernot et al., 2017)	40
17	Dwa histogramy obrazujące rozkład liczby znaków wiadomości w zależności od liczby znaków dla treści spam oraz tekstów zaufanych(Kuchipudi et al., 2020)	44

Spis tabel

1	Interpretacja oceny punktowej BIK(scoringexpert.pl, 2017)	14
2	Klasyczna karta oceny punktowej(Przanowski, 2023)	19
3	Klasyczna karta oceny punktowej(Feng B., 2021)	38
4	Wrogie próbki zastosowane na atakowanym modelu filtra antyspamowego (Kuchipudi et al., 2020)	42

Załączniki