Politechnika Wrocławska

Wydział Matematyki
Kierunek studiów: Matematyka stosowana
Specjalność: –

Praca dyplomowa – inżynierska

WYKRYWANIE OSZUSTW NA KARTACH PŁATNICZYCH Z WYKORZYSTANIEM METOD WRAŻLIWYCH NA KOSZT

Patryk Wielopolski

Słowa kluczowe:
Uczenie maszynowe; Klasyfikacja wrażliwa na koszt; Wykrywanie oszustw na kartach kredytowych

Krótkie streszczenie:
Celem pracy była implementacja eksperymentu, który pozwoli porównać klasyczne oraz wrażliwe na koszt metody predykcyjne w problemie detekcji oszustw na kartach płatniczych. W pracy rozważono dwa rodzaje podejść do modelowania wrażliwego na koszt – klasyfikację wrażliwą na koszt oraz trening wrażliwy na koszt. Na podstawie wyników eksperymentów udało się pokazać, że takie podejście do modelowania daje znacznie lepsze rezultaty niż standradowe metody. Ponadto pokazaliśmy, że niewykorzysytwany do tej pory w takich problemach algorytm XGBoost w połączeniu z minimalizacją ryzyka bayesowskiego uzyskuje lepsze rezultaty niż wcześniej wykorzysytwane metody.

Opiekun pracy dyplomowej	dr inż. Andrzej Giniewicz	.....	.....
	Tytuł/stopień naukowy/imię i nazwisko	ocena	podpis

*Do celów archiwalnych pracę dyplomową zakwalifikowano do:**

a) kategorii A (akta wieczyste)

b) kategorii BE 50 (po 50 latach podlegające ekspertyzie)

** niepotrzebne skreślić*

pieczątka wydziałowa

Wrocław, rok 2020

Wrocław University
of Science and Technology

Faculty of Pure and Applied Mathematics
Field of study: Applied Mathematics
Specialty: –

Engineering Thesis

CREDIT CARD FRAUD DETECTION USING COST SENSITIVE METHODS

Patryk Wielopolski

Keywords:
Machine Learning; Cost Sensitive Classification; Credit Card Fraud Detection

Short summary:
The aim of this thesis was an implementation of an experiment, which will allow us to compare classical and cost-sensitive methods in credit card fraud detection. The study considers two types of cost-sensitive modeling approaches – cost-sensitive classification and cost-sensitive training. Based on the results of the experiments, it was shown that this modeling approach gives much better results than standard methods. In addition, we have shown that the XGBoost algorithm not used so far in such problems in combination with Bayesian Minimum Risk achieves better results than the previously used methods.

Supervisor	dr inż. Andrzej Giniewicz	.....	.....
	Title/degree/name and surname	grade	signature

*For the purposes of archival thesis qualified to:**

a) category A (perpetual files)

b) category BE 50 (subject to expertise after 50 years)

** delete as appropriate*

stamp of the faculty

Wrocław, 2020

Spis treści

Wstęp **1**

1 Wprowadzenie teoretyczne **3**
1.1 Miary skuteczności modeli 4
1.1.1 Macierz pomyłek 4
1.1.2 Miary skuteczności niewrażliwe na koszt 4
1.1.3 Krzywa ROC 7
1.1.4 Macierz kosztu 8
1.1.5 Miary skuteczności modeli wrażliwych na koszt 9
1.2 Standardowe modele 10
1.2.1 Regresja logistyczna 10
1.2.2 Drzewo decyzyjne 12
1.2.3 Las losowy 14
1.2.4 XGBoost 15
1.3 Klasyfikacja wrażliwa na koszt 17
1.3.1 Optymalizacja progu 17
1.3.2 Minimalizacja ryzyka bayesowskiego 18
1.4 Trening wrażliwy na koszt 19
1.4.1 Regresja logistyczna wrażliwa na koszt 20
1.4.2 Drzewo decyzyjne wrażliwe na koszt 21

2 Eksperyment **23**
2.1 Dane 23
2.2 Opis eksperymentu 25
2.3 Wyniki 25

Podsumowanie **29**

A Minimalizacja ryzyka bayesowskiego **31**

Bibliografia **33**

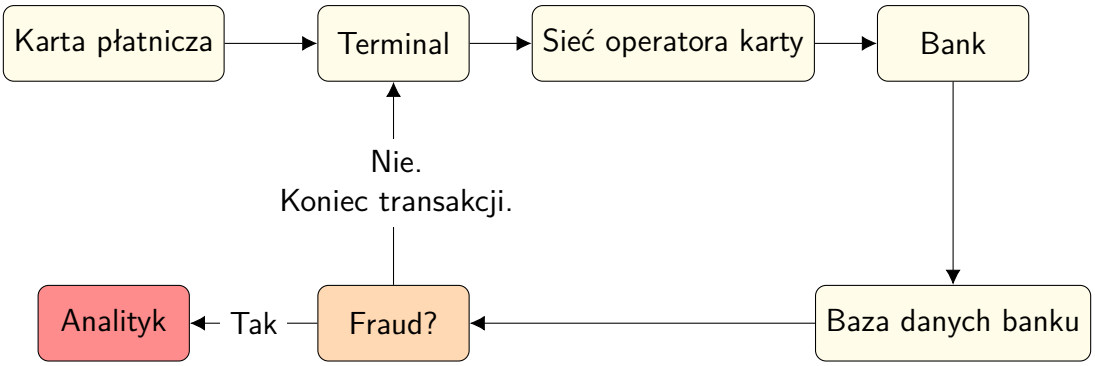
Wstęp

Historia kart płatniczych zaczyna się już w latach 80. XIX wieku, kiedy Edward Bellamy w swojej powieści *Looking Backward* wspomina o możliwości skorzystania z karty przedpłaconej (karta umożliwiająca dokonanie płatności z wcześniej wpłaconych środków)[1]. Niestety pomysł takiej formy płatności nie został pozytywnie przyjęty w tamtych czasach i dopiero w 1914 roku amerykańska firma Western Union umożliwiła korzystanie z kart obciążeniowych (bank udziela klientowi odpowiedniego limitu płatności, za które należy zapłacić później). Koncepcja karty bardzo szybko rozpowszechniła się i mimo wybuchu II Wojny Światowej, która wstrzymała jej rozwój, to już w 1950 roku firma Diners Club umożliwiła swoim klientom dokonywania płatności w sklepach, restauracjach lub stacjach benzynowych bez użycia gotówki[2]. Kilka lat później Bank of America, wydał pierwszą prawdziwą kartę kredytową, która umożliwiała spłatę zadłużenia w całości lub tylko pewnej wyliczonej kwoty minimalnej. Kolejnym przełomem było wprowadzenie w roku 1972 paska magnetycznego, który umożliwiał korzystanie z PINu i dzięki temu zwiększał bezpieczeństwo klientów[3]. W latach 1973 i 1974 wprowadzono odpowiednio system BASE I oraz BASE II, które były pierwszymi systemami elektronicznymi. Następnie w latach dziewięćdziesiątych pojawiają się technologie, które mają zapewnić zwiększenie bezpieczeństwa oraz zmniejszenie skali oszustw.

Dalszy rozwój technologiczny umożliwił nie tylko dokonywanie płatności kartą, lecz także wykorzystywanie do tego celu telefonu komórkowego lub zegarka. Ponadto transakcji nie dokonujemy już tylko w sklepach stacjonarnych, lecz także coraz częściej korzystamy ze sklepów internetowych. Wszystkie te czynniki powodują, że dokonywanie zakupów jest jeszcze prostsze, a ilość dokonywanych transakcji z roku na rok rośnie[4]. Niestety ten wzrost spowodował również rozwój sposobów dokonywania przestępstw związanych z kartami płatniczymi, przykładowo przekazanie numeru karty nieznajomym, utrata lub kradzież karty, skopiowania danych karty lub kradzież przesyłki z kartą, które najczęściej kończą się nieautoryzowanymi płatności na koncie.

Banki oraz instytucje finansowe w celu przeciwdziałania tym przestępstwom korzystają z systemów detekcji oszustw. Jest to jeden z elementów całego procesu dokonywania płatności, który jest przedstawiony na rysunku 1 na stronie 2. W momencie przyłożenia karty kredytowej do terminala nawiązywane jest połączenie z operatorem karty (np. Visa, MasterCard), który dalej komunikuje się z bankiem klienta. Następnie w odpowiednich bazach danych sprawdzane są dostępne środki, a system detekcji oszustw sprawdza, czy transakcja nie jest podejrzana. W przypadku wykrycia niezgodności system ją odrzuca i kieruje sprawę do odpowiedniego analityka. W przeciwnym razie płatność jest akceptowana i na terminalu wyświetla się odpowiedni komunikat.

[1] Źródło: `https://pl.wikipedia.org/wiki/Karta_p%C5%82atnicza` (dostęp: 27.12.2019)
[2] Źródło: `https://www.kartyonline.pl/historia-kart-platniczych.php` (dostęp: 27.12.2019)
[3] Źródło: `https://en.wikipedia.org/wiki/Payment_card` (dostęp: 27.12.2019)
[4] Źródło: `https://worldpaymentsreport.com/non-cash-payments-volume/` (dostęp: 27.12.2019)



Rysunek 1: Schemat płatności kartą płatniczą. Źródło: Opracowanie własne na podstawie prezentacji A. Bahnsena.

Do niedawna najbardziej popularnym sposobem do wykrywania oszustw był system oparte na regułach eksperckich[5]. Powstały one na bazie doświadczenia analityków, którzy podczas swojej długoletniej pracy wielokrotnie sprawdzali transakcje pod kątem oszustw i byli w stanie dostrzec pewne zależności. Przykładami takich reguł mogą być: więcej niż cztery wypłaty z bankomatu w ciągu godziny, więcej niż dwie transakcje w ciągu 5 minut, dwie transakcje w ciągu 5 minut różnymi formami płatności (karta, internet). W przypadku, gdy rozpatrywana transakcja spełnia chociaż jedną z reguł, to jest ona blokowana i użytkownik jest informowany o odrzuceniu transakcji. To podejście jest stosunkowo łatwe do implementacji oraz jest bardzo proste do interpretacji, lecz niestety ma swoje wady. Oszuści często wraz z upływem czasu zmieniają swoje sposoby dokonywania oszustw, natomiast systemy złożone z reguł eksperckich nie są w stanie same wykrywać nowych zachowań i dostosowywać się do zmian, lecz wymagają do tego dodatkowej pracy analityków. Tę niedogodność są w stanie rozwiązać modele uczenia maszynowego, które tworzą pewnego rodzaju reguły, które nie są bezpośrednio wprowadzane przez użytkownika, lecz automatycznie wykrywane przez algorytmy na podstawie dostarczonych danych historycznych. Wraz z dynamicznym rozwojem tej dziedziny nauki zaczęły powstawać coraz bardziej wyrafinowane modele, które coraz lepiej radzą sobie z wykrywaniem takich zależności. Obecnie są one już wykorzystywane w produkcyjnych środowiskach instytucji finansowych i wspomagają walkę z przestępcami. Niestety standardowe podejście do modelowania predykcyjnego z wykorzystaniem uczenia maszynowego zakłada równy koszt popełnienia błędu, co nie jest założeniem odpowiadającym rzeczywistości. Stąd powstała motywacja do rozwoju metod, które biorą pod uwagę tę różnicę.

Celem pracy jest implementacja eksperymentu, który pozwoli porównać standardowe oraz wrażliwe na koszt metody predykcyjne w problemie detekcji oszustw na kartach płatniczych. Rozpoczniemy od wstępu teoretycznego w rozdziale pierwszym, który wprowadzi odpowiednie pojęcia i modele z dziedziny statystyki matematycznej oraz uczenia maszynowego. Następnie w rozdziale drugim opiszemy przeprowadzone doświadczenie na danych rzeczywistych, w którym zostały wykorzystane modele takie jak regresja logistyczna, drzewo decyzyjne, las losowy i algorytm XGBoost wraz z odpowiednimi modyfikacjami oraz metody optymalizacji progu i minimalizacji ryzyka bayesowskiego. W dalszej kolejności omówimy otrzymane rezultaty oraz wyciągniemy wnioski. Finalnie w ostatniej części zawrzemy podsumowanie.

[5]Wystąpienie A.Bahnsena podczas Konferencji Analytics 2013. Źródło: `https://www.youtube.com/watch?v=YCNkxaVDiA0`

Rozdział 1

Wprowadzenie teoretyczne

Wprowadzenie teoretyczne zawiera przegląd metod z dziedziny statystyki oraz uczenia maszynowego, które wykorzystuje się w zagadnieniach klasyfikacyjnych. W szczególności zostanie opisana macierz pomyłek, macierz kosztu oraz związane z nimi miary skuteczności modeli predykcyjnych, które służą do badania jakości predykcji. Następnie zostaną omówione modele standardowych, które nie biorą pod uwagi różnych kosztów popełniania błędu klasyfikacyjnego. W ramach tej grupy zostanie opisana regresja logistyczną, drzewo decyzyjne, las losowy oraz algorytm XGBoost. Na końcu zostanie poruszony temat modeli wrażliwych na koszt, które dzielimy na dwie podkategorie: trening wrażliwy na koszt (*ang. Cost Sensitive Training*) oraz klasyfikacja wrażliwa na koszt (*ang. Cost Sensitive Classification*)[14]. Pierwsza z nich zawiera metody, które w trakcie uczenia modelu zwracają uwagę na koszt błędnej klasyfikacji. Są to rozszerzone wersje standardowych modeli, które w trakcie uczenia się mają zmienioną postać optymalizowanej funkcji. W niniejszej pracy będziemy omawiać regresję logistyczną wrażliwą na koszt oraz drzewo decyzyjne wrażliwe na koszt. Druga podkategoria składa się z metod, które wykorzystują estymowane prawdopodobieństwo ze standardowych modeli i na jego podstawie tworzą dodatkowy model, który bierze pod uwagę koszt niepoprawnej klasyfikacji danego przypadku. Z tej grupy zostanie omówiona optymalizacja progu oraz minimalizacja ryzyka bayesowskiego.

W pracy będziemy wykorzystywać następujące oznaczenia:

- $\mathcal{S} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ – zbiór wszystkich N obserwacji,
- $\boldsymbol{x}_i = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(j)}\right)$ – i-ty poziomy wektor obserwacji składający się z j atrybutów,
- $x_i^{(k)}$ – k-ty atrybut i-tego wektora obserwacji,
- $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ - poziomy wektor prawdziwych stanów klasyfikacji,
- $y_i \in \{0, 1\}$ – prawdziwy stan klasyfikacji dla i-tej obserwacji,
- $p_i \in [0, 1]$ – estymowana wartość prawdopodobieństwa dla i-tej obserwacji,
- $\boldsymbol{c} = (c_1, c_2, \ldots, c_N)$ - poziomy wektor przewidywanych klas
- $c_i \in \{0, 1\}$ – przewidywana klasa dla i-tej obserwacji.

Ponadto wartość 0 zmiennej y_i oraz c_i oznacza klasyfikację negatywną, natomiast wartość 1 oznacza klasyfikację pozytywną.

1.1 Miary skuteczności modeli

1.1.1 Macierz pomyłek

Jedną z metod wykorzystywanych do oceny skuteczności modeli w zadaniach klasyfikacyjnych jest macierz pomyłek. Prawidłowe stany zestawiamy z klasami nadanymi przez model, a następnie sprawdzamy, czy poprawnie zaklasyfikowaliśmy poszczególne obserwacje. W przypadku klasyfikacji binarnej popełniamy dwa rodzaje błędów, a wszystkie możliwe wyniki prezentują się w następujący sposób:

– Prawdziwie pozytywny wynik klasyfikacji (TP, z angielskiego *True Positive*) – obserwacja miała stan sprzyjający i zaklasyfikowaliśmy ją jako pozytywną;

– Fałszywie pozytywny wynik klasyfikacji (FP, z angielskiego *False Positive*) – obserwacja miała stan niesprzyjający i zaklasyfikowaliśmy ją jako pozytywną. W nomenklaturze statystycznej nazywamy ją również błędem pierwszego rodzaju;

– Fałszywie negatywny wynik klasyfikacji (FN, z angielskiego *False Negative*) – obserwacja miała stan sprzyjający i zaklasyfikowaliśmy ją jako negatywną. W nomenklaturze statystycznej nazywamy ją również błędem drugiego rodzaju;

– Prawdziwie negatywny wynik klasyfikacji (TN, z angielskiego *True Negative*) – obserwacja miała stan niesprzyjający i zaklasyfikowaliśmy ją jako negatywną.

Liczbę obserwacji należących do poszczególnych kategorii możemy reprezentować tak jak w tabeli 1.1, gdzie

- $\mathrm{TP} = \#\{i : i \in \{1, \ldots, N\}, y_i = c_i = 1\}$,
- $\mathrm{FP} = \#\{i : i \in \{1, \ldots, N\}, y_i = 0, c_i = 1\}$,
- $\mathrm{FN} = \#\{i : i \in \{1, \ldots, N\}, y_i = 1, c_i = 0\}$,
- $\mathrm{TN} = \#\{i : i \in \{1, \ldots, N\}, y_i = c_i = 0\}$.

	Stan sprzyjający $y_i = 1$	Stan niesprzyjający $y_i = 0$
Predykcja pozytywna $c_i = 1$	TP	FP
Predykcja negatywna $c_i = 0$	FN	TN

Tabela 1.1: Macierz pomyłek.

1.1.2 Miary skuteczności niewrażliwe na koszt

Na podstawie macierzy pomyłek (tabela 1.1) możemy zdefiniować następujące miary skuteczności modeli, które nie biorą pod uwagę kosztu popełnienia błędu. Przytoczone definicje pochodzą z pracy D.A. Powersa [15].

Dokładność

Dokładność (*ang. Accuracy*) mierzy stosunek poprawnie zaklasyfikowanych przypadków do ilości wszystkich obserwacji. Bierze ona pod uwagę zarówno obserwacje pozytywne, jak i negatywne. Z tego powodu bardzo dobrze radzi sobie w sytuacji, gdy interesują nas obserwacje z obu klas, natomiast dużo gorzej, gdy istotna jest tylko jedna klasa. W szczególności nie nadaje się ona do problemów z dużą dysproporcją klas, która pojawia się w przypadku detekcji oszustw. Dokładność zdefiniowana jest następującym wzorem

$$\text{Dokładność} = \frac{TP + TN}{TP + FP + FN + TN} = \alpha \cdot \text{Czułość} + (1 - \alpha) \cdot \text{Swoistość},$$

gdzie α to odsetek przypadków pozytywnych w danych, a czułość oraz swoistość są opisane w następnych podsekcjach.

W celu uniknięcia różnic w obu populacjach wykorzystuje się zbalansowaną skuteczność (*BACC, ang. balanced accuracy*), która za parametr α przyjmuje wartość $\frac{1}{2}$. W przypadku, gdy klasyfikator działa równie dobrze dla obu klas, to sprowadza się ona do dokładności, natomiast gdy standardowa dokładność jest wysoka tylko z uwagi na niezbalansowaną liczność próbek w klasach, to wartość zbalansowanej skuteczności spada [8]. Jest ona również ważna, ponieważ maksymalizacja tej miary odpowiada standardowemu szukaniu optymalnego punktu na krzywej ROC (patrz sekcja 1.1.3 na stronie 7) wg. metryki taksówkowej.

Precyzja

Precyzja (*ang. Precision*) to stosunek poprawnie zaklasyfikowanych pozytywnych przypadków do wszystkich obserwacji zaklasyfikowanych jako pozytywne. Mierzy ona, jaki procent próbek prawidłowo zaklasyfikowaliśmy jako pozytywne. Warto zauważyć, że nie bierze ona pod uwagę obserwacji negatywnych. Jest ona bardzo często wykorzystywana podczas rozwiązywania większości problemów z zakresu uczenia maszynowego, ponieważ najczęściej interesuje nas, aby nasze typowania były możliwie najbardziej skuteczne. Z tego też powodu precyzja jest często nazywana skutecznością klasyfikacji prawdziwie pozytywnych (*ang. True Positive Accuracy*). Zapisana wzorem przyjmuje postać

$$\text{Precyzja} = \frac{TP}{TP + FP}.$$

Rozpatrzmy jeszcze przykład z medycyny. Załóżmy, że mamy test diagnostyczny, który klasyfikuje wszystkich pacjentów jako chorych. W tym przypadku czułość przyjmuje wartość 1, swoistość wartość 0, a precyzja jest równa frakcji przypadków pozytywnych w danych. Odwrotnie, załóżmy teraz, że nasz test oznacza wszystkich pacjentów jako zdrowych. W tym przypadku czułość jest równa 0, swoistość 1, natomiast precyzja jest niezdefiniowana (symbol nieoznaczony – $\frac{0}{0}$). Ten przykład pokazuje, że precyzja jest zależna od zbalansowania klas.

Czułość

Czułość (*ang. Recall*) to stosunek prawdziwie pozytywnie zaklasyfikowanych przypadków do wszystkich pozytywnych obserwacji. Innymi słowy, jest to miara, która mówi nam o procencie znalezionych przez model przypadków pozytywnych. Podobnie jak precyzja skupia się jedynie na poprawnie zaklasyfikowanych obserwacjach pozytywnych. Jest ona

szczególnie ważna w zastosowaniach medycznych, gdzie naszym głównym celem jest znalezienie wszystkich chorych osób. Używana jest także do wyznaczania krzywej ROC opisanej w sekcji 1.1.3 na stronie 7. Czułość jest wyrażona wzorem

$$\text{Czułość} = \frac{TP}{TP + FN}.$$

Swoistość

Swoistość (*ang. Specificity*) to stosunek prawdziwie negatywnie zaklasyfikowanych przypadków do wszystkich negatywnych obserwacji. Inaczej nazywana też odwrotną czułością, z uwagi na analogiczną postać do czułości, ponieważ podobnie mierzy ona odsetek znalezionych wszystkich obserwacji, lecz dla negatywnej klasy. Wyraża się następującym wzorem

$$\text{Swoistość} = \frac{TN}{TN + FP}.$$

Ujemna wartość predykcyjna

Ujemna wartość predykcyjna to stosunek prawdziwie negatywnie zaklasyfikowanych przypadków do wszystkich negatywnie zaklasyfikowanych obserwacji. Nazywana jest również odwrotną precyzją, ponieważ analogicznie do precyzji skupia się na mierzeniu skuteczności klasyfikowania obserwacji negatywnych. Miara ta jest także używana do wyrysowania krzywej ROC (patrz sekcja 1.1.3 na stronie 7), a dokładniej wartość $1 -$ ujemna wartość predykcyjna, która odpowiada stosunekowi fałszywie pozytywnych klasyfikacji (ang. *false positive ratio*). Dana jest wzorem

$$\text{Ujemna wartość predykcyjna} = \frac{TN}{TN + FN}.$$

F Score

F_1 Score to miara łącząca w sobie precyzję oraz czułość za pomocą średniej harmonicznej. Z uwagi na wykorzystanie tych dwóch wartości nie bierze ona pod uwagę prawidłowo zaklasyfikowanych obserwacji negatywnych. Wykorzystanie średniej harmonicznej powoduje, że kładziemy równy nacisk na obie użyte miary. Standardowy wzór wygląda następująco

$$\text{F}_1 \text{ Score} = \left(\frac{2}{\text{Precyzja}^{-1} + \text{Czułość}^{-1}} \right) = 2 \cdot \frac{\text{Precyzja} \cdot \text{Czułość}}{\text{Precyzja} + \text{Czułość}}.$$

W przypadku, gdy interesuje nas bardziej precyzja lub czułość, to definiujemy bardziej ogólną formułę na F_β Score:

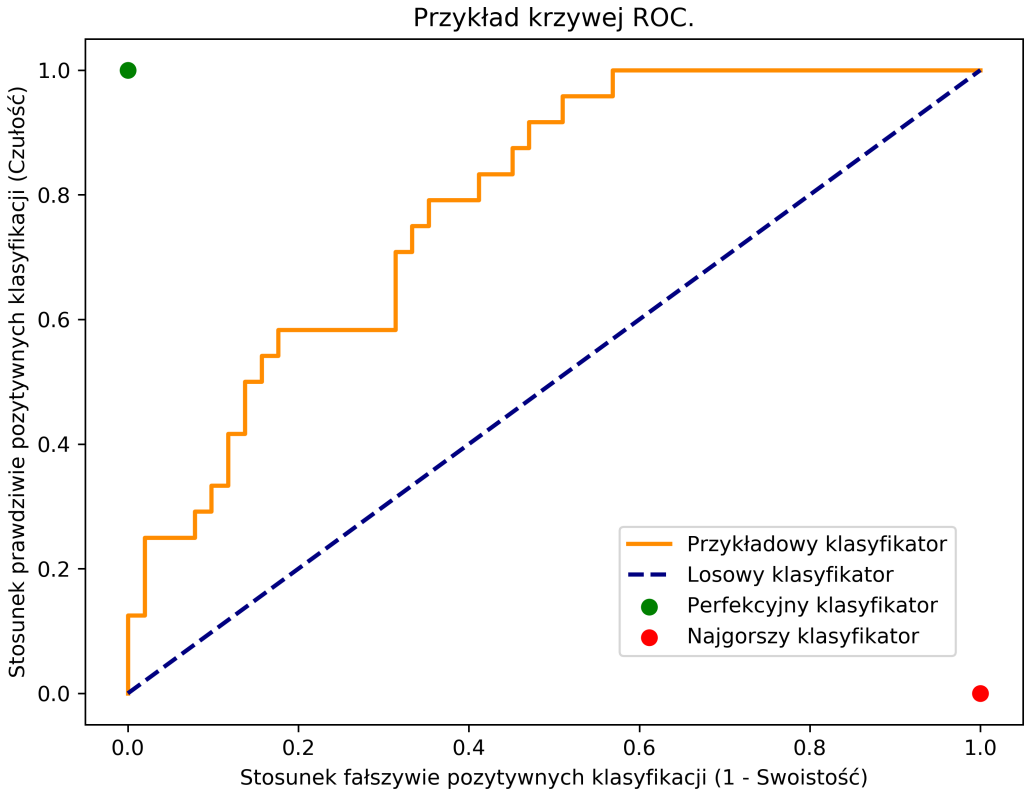
$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precyzja} \cdot \text{Czułość}}{(\beta^2 \cdot \text{Czułość}) + \text{Precyzja}},$$

gdzie parametr $\beta > 0$ odpowiada na pytanie, ile razy bardziej czułość jest ważna niż precyzja. Najczęściej wykorzystuje się parametr $\beta = 2$ lub $\beta = \frac{1}{2}$, który odpowiednio zwiększa nacisk na czułość lub zmniejsza wpływ fałszywie negatywnie zaklasyfikowanych obserwacji, skupiając się bardziej na precyzji.

1.1.3 Krzywa ROC

Krzywa ROC to wykres diagnostyczny modelu predykcyjnego pozwalający na zilustrowanie zdolności klasyfikatora do rozróżniania poszczególnych klas. Składa się ona z wyrysowanym na osi X stosunku fałszywie pozytywnych klasyfikacji na oraz czułości na osi Y [15]. Miary są narysowane dla różnych wartości progu decyzyjnego, który prawdopodobieństwa zwracane przez model zamienia na decyzje binarne. Perfekcyjny klasyfikator jest w stanie osiągnąć lewy górny róg wykresu, czyli posiadać precyzję na poziomie 100% oraz stosunek fałszywie pozytywnych klasyfikacji na poziomie 0%. Analogicznie najgorszy klasyfikator jest w stanie osiągnąć prawy dolny róg wykresu. Modelem losowy będzie punkt na prostej z punktu $(0, 0)$ do punktu $(1, 1)$.

Przykładowy wykres znajduje się na rysunku 1.1 na stronie 7. Kolorem czerwonym oznaczony jest najgorszy możliwy klasyfikator, zielonym perfekcyjny, niebieskim losowy, natomiast żółtym przykładowy.



Rysunek 1.1: Przykładowy wykres charakterystyki pracy odbiornika. Na poziomej osi przedstawiony jest stosunek fałszywie pozytywnych klasyfikacji, natomiast na pionowej czułość. Kolorem czerwonym, zielonym, niebieskim oraz żółtym jest odpowiednio oznaczony najgorszy, najlepszy, losowy oraz przykładowy klasyfikator. Źródło: Opracowanie własne.

Celem tworzenia oraz ulepszania kolejnych modeli predykcyjnych jest zbliżanie się w wynikach do lewego górnego rogu wykresu. Odległość danego punktu reprezentującego konkretny model możemy definiować w różnych metrykach. Przykładowo w metryce taksówkowej (metryka L_1) będzie to

$$1 - \text{Swoistość} + 1 - \text{Czułość}.$$

Dokonując kolejne przekształcenia otrzymujemy

$$1 - \text{Swoistość} + 1 - \text{Czułość} = 2 - (\text{Czułość} + \text{Swoistość}) =$$
$$= 2 \cdot (1 - (\text{Czułość} + \text{Swoistość})/2) = 2 * (1 - \text{BACC}).$$

Oznacza to, że im większa zbalansowana skuteczność danego modelu, tym bliżej znajduje się model na krzywej ROC lewego górnego rogu. W podobny sposób możemy także rozważać inne metryki. Przykładowo w metryce euklidesowej (metryka L_2)będziemy chcieli minimalizować następujące wyrażenie

$$\sqrt{(1 - \text{Swoistość})^2 + (1 - \text{Czułość})^2}$$

bądź ogólniej w metryce L_p

$$\Big((1 - \text{Swoistość})^p + (1 - \text{Czułość})^p\Big)^{\frac{1}{p}}.$$

Z uwagi na fakt, że w przypadku metryki L_1 uzyskaliśmy bezpośrednią miarę skuteczności do optymalizacji, to istnieje szansa, że dla metryki L_2 oraz ogólnie L_p istnieją również analogiczne miary. Natomiast w toku przeprowadzanej pracy nie udało się uzyskać takiego rezultatu.

1.1.4 Macierz kosztu

W celu wprowadzenia potrzebnych miar skuteczności wrażliwych na koszt potrzebujemy najpierw zdefiniować wektor macierzy kosztu

$$\boldsymbol{C} = (C_1, C_2, \ldots, C_N)$$

gdzie $C^{(i)} = [C^{(i)}_{j,k}]^1_{j,k=0}$ oznacza macierz kosztu dla i-tej obserwacji [5, 12]. Ma ona postać analogiczną do macierzy pomyłek, lecz zamiast ilość obserwacji w odpowiednich komórkach wpisujemy koszt pomyłki dla i-tej obserwacji. Warto zaznaczyć, że koszt niekoniecznie musi być kwota pieniężną, lecz także możemy rozważać ilość poświęconego czasu, bądź inną dowolną mierzalną jednostkę. W szczególności dla przypadku klasyfikacji binarnej wyróżniamy cztery możliwe wartości:

- $C_{1,1}$ – koszt prawdziwie pozytywnego zaklasyfikowania i-tej obserwacji,
- $C_{1,0}$ – koszt fałszywie pozytywnego zaklasyfikowania i-tej obserwacji,
- $C_{0,1}$ – koszt fałszywie negatywnego zaklasyfikowania i-tej obserwacji,
- $C_{0,0}$ – koszt prawdziwie negatywnego zaklasyfikowania i-tej obserwacji.

W zależności od praktycznego zastosowania wartości w macierzy kosztu przyjmują różne wielkości. Poniżej przedstawiamy kilka intuicji na podstawie macierzy kosztu zdefiniowanych dla różnych eksperymentów w pracy A. Bahnsena [2].

- W przypadku, gdy predykcja klasy pozytywnej wiąże się z podjęciem jakiejś akcji, np. sprawdzenie transakcji kartą kredytową, wysłanie oferty reklamowej, to koszt pozytywnej predykcji jest dodatni, to znaczy $C^{(i)}_{1,1}$, $C^{(i)}_{1,0} > 0$. Ponadto najczęściej te koszty są równe i przyjmują stałą wartość dla wszystkich przypadków, to znaczy $C^{(i)}_{1,1} = C^{(i)}_{1,0} = \text{const.}$, ponieważ podejmowana zawsze jest ta sama akcja dla wszystkich

obserwacji. Dodatkowo koszt pozytywnego zaklasyfikowania dla każdej obserwacji negatywnej jest zerowy, to znaczy $C^{(i)}_{0,0} = 0$, ponieważ słusznie nie została podjęta żadna akcja. Zatem ostatecznie jedyną wartością, która różni się między przypadkami, jest koszt fałszywie negatywnej klasyfikacji i on również jest większy od zera.

- W przypadku, gdy predykcja dokonywana jest przez zautomatyzowany system, to koszt prawidłowej klasyfikacji jest równy zero, to znaczy $C^{(i)}_{1,1} = C^{(i)}_{0,0} = 0$, ponieważ maszynie nie musimy płacić pensji, która obsługuje daną sprawę. Natomiast koszty pomyłek są różne i zazwyczaj różnią się między obserwacjami. Biorąc jako przykład ryzyko kredytowe, możemy zauważyć, że koszty mogą być zarówno dodatnie, jak i ujemne. W tym konkretnym przypadku, gdy stwierdzimy, że dana osoba spłaci kredyt, natomiast w rzeczywistości nie zrobi tego, to będziemy stratni pewną sumę pieniędzy, co da nam dodatni koszt. Z drugiej strony, gdy uznamy, że dana osoba nie spłaci kredytu, a w rzeczywistości wywiązałaby się z zobowiązań, to utracimy pewien zarobek, zatem koszt będzie negatywny.

Koncepcyjnie koszt popełnienia błędu dla każdej obserwacji powinien być większy niż koszt poprawnej klasyfikacji, to znaczy

$$C^{(i)}_{1,0} > C^{(i)}_{1,1} \text{ i } C^{(i)}_{0,1} > C^{(i)}_{0,0} \quad \forall_{i \in \{1,2,\ldots,N\}}.$$

Natomiast jak możemy zauważyć na podstawie wyżej opisanych intuicji, w praktycznych zastosowaniach jest to bardzo często nierealistyczne założenie. Przykład takiej macierzy kosztu znajduje się w tabeli 1.2.

	Stan pozytywny $y_i = 1$	Stan negatywny $y_i = 0$
Predykcja pozytywna $c_i = 1$	$C^{(i)}_{1,1}$	$C^{(i)}_{1,0}$
Predykcja negatywna $c_i = 0$	$C^{(i)}_{0,1}$	$C^{(i)}_{0,0}$

Tabela 1.2: Macierz kosztu dla i-tej obserwacji.

1.1.5 Miary skuteczności modeli wrażliwych na koszt

Motywacją do powstania miar skuteczności wrażliwych na koszt jest potrzeba ewaluacji modeli, która miarodajnie odda złożoność rozwiązywanego problemu [1]. Przykładowo w przypadku detekcji oszustw będzie w stanie odpowiedzieć na pytanie dotyczące zaoszczędzonej kwoty dzięki modelowi predykcyjnemu. Kolejny powodem jest fakt, że podstawowe metryki nie uwzględniają różnicy w kosztach pomyłki dla fałszywie pozytywnych oraz fałszywie negatywnych klasyfikacji, traktując oba błędy jednakowo. Ponadto warto zauważyć, że omawiane problemy wymagają uwzględnienia nie tylko różnych kosztów per rodzaj pomyłki, ale także per konkretna obserwacja. Bazując na wprowadzonym w sekcji 1.1.4 wektorze macierzy kosztu C, zdefiniujemy kilka miar.

Koszt całkowity

Koszt całkowity (TC, z angielskiego *ang. Total Cost*) to suma kosztów poniesionych dla danego zestawy predykcji $\boldsymbol{c}$, który zostały wygenerowany przez pewien model predykcyjny. Definiowany jest wzorem

$$\mathrm{TC}(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{C}) = \sum_{i=1}^{N} C^{(i)}_{c_i, y_i}, \tag{1.1}$$

Oszczędności

Oszczędności (*ang. Savings*) to miara, którą intuicyjnie możemy rozumieć jako procentową wartość, o ile testowany model jest lepszy od naiwnego klasyfikatora. Przy czym naiwnym klasyfikatorem nazywamy model, który zwraca same predykcje pozytywne lub same predykcje negatywne, w zależności od tego, co bardziej się opłaca dla danego zestawu danych.

$$\text{Oszczędności}(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{C}) = \frac{\text{Koszt bazowy}(\boldsymbol{y}, \boldsymbol{C}) - \mathrm{TC}(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{C})}{\text{Koszt bazowy}(\boldsymbol{y}, \boldsymbol{C})}, \tag{1.2}$$

gdzie:

- $\text{Koszt bazowy}(\boldsymbol{y}, \boldsymbol{C}) = \min\{\mathrm{TC}(\boldsymbol{y}, \boldsymbol{c}_0, \boldsymbol{C}), \mathrm{TC}(\boldsymbol{y}, \boldsymbol{c}_1, \boldsymbol{C})\} \neq 0$ – koszt poniesiony w przypadku używania naiwnego klasyfikatora,
- $\boldsymbol{c}_0 = (0, 0, \ldots, 0)$ – N-elementowy wektor poziomy predykcji równych 0,
- $\boldsymbol{c}_1 = (1, 1, \ldots, 1)$ – N-elementowy wektor poziomy predykcji równych 1.

1.2 Standardowe modele

1.2.1 Regresja logistyczna

Regresja logistyczna to model statystyczny, którego celem jest modelowanie prawdopodobieństwa, że zmienna losowa Y przyjmie wartość 0 lub 1, na podstawie danych empirycznych. Rozważmy następującą definicję modelu regresji logistycznej parametryzowaną poziomym wektorem współczynników $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ daną w następujący sposób

$$h_\theta(\boldsymbol{x_i}) = \frac{1}{1 + e^{-\boldsymbol{x}_i \theta^T}} = P(Y = 1 | X = \boldsymbol{x}_i; \theta). \tag{1.3}$$

Jednocześnie zauważmy, że

$$P(Y = 0 | X = \boldsymbol{x}_i; \theta) = 1 - P(Y = 1 | X = \boldsymbol{x}_i; \theta) = 1 - h_\theta(\boldsymbol{x_i})$$

oraz korzystając z faktu, że $y_i \in \{0, 1\}$, możemy zapisać

$$P(Y = y_i | \boldsymbol{x}_i; \theta) = h_\theta(\boldsymbol{x_i})^{y_i} (1 - h_\theta(\boldsymbol{x_i}))^{(1 - y_i)}.$$

Następnie wyznaczamy funkcję największej wiarygodności, aby wyestymować parametry θ.

$$L(\theta | \mathcal{S}, \boldsymbol{y}) = \prod_{i=1}^{N} P(Y = y_i | X = \boldsymbol{x}_i; \theta) = \prod_{i=1}^{N} h_\theta(\boldsymbol{x_i})^{y_i} (1 - h_\theta(\boldsymbol{x_i}))^{(1 - y_i)}$$

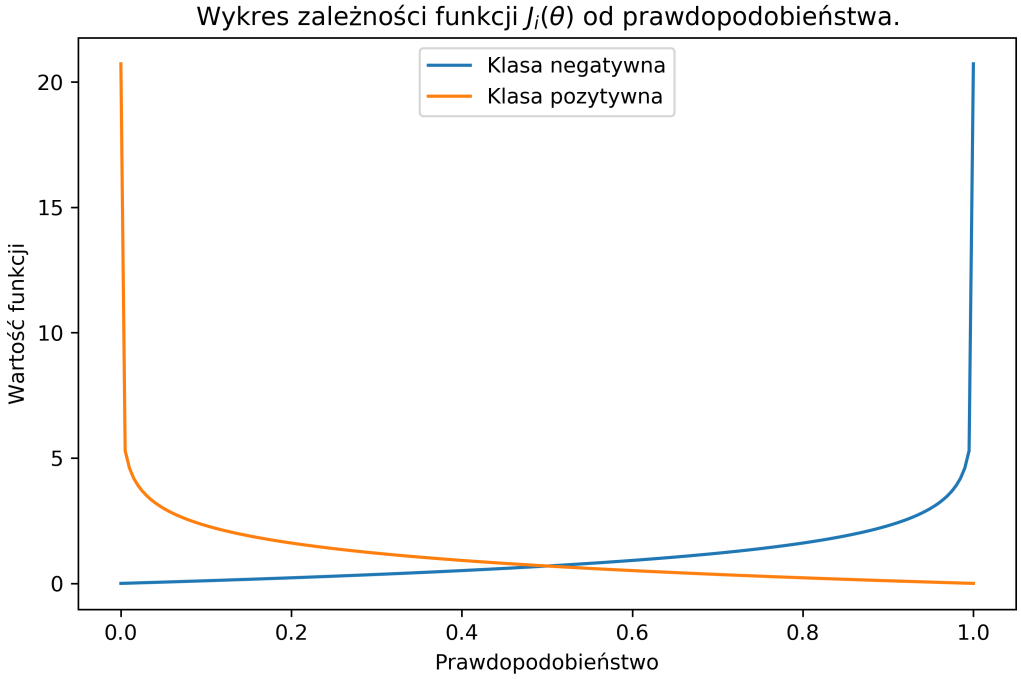
W celu znalezienia maksimum funkcji wiarygodności logarytmujemy funkcję $L(\theta|\mathcal{S}, \boldsymbol{y})$ i otrzymujemy funkcję straty $J(\theta)$

$$J(\theta) = \log L(\theta|\mathcal{S}, \boldsymbol{y}) = -\sum_{i=1}^{N} J_i(\theta),$$

gdzie $J_i(\theta)$ nazywana jest entropią krzyżową i przyjmuje postać

$$J_i(\theta) = -\Big(y_i \log(h_\theta(\boldsymbol{x_i})) + (1 - y_i) \log(1 - h_\theta(\boldsymbol{x_i}))\Big). \tag{1.4}$$

Niestety nie jest możliwe znalezienie bezpośredniego wzoru na współczynniki, które maksymalizują funkcję wiarygodności, zatem wykorzystuje się w tym celu algorytmy optymalizacji numerycznej, np. IRLS (*ang. Iteratively Reweighted Least Squares*).



Rysunek 1.2: Wykres zależności funkcji $J_i(\theta)$ od prawdopodobieństwa predykcji danej obserwacji. Kolor niebieski przedstawia wykres dla próbki o prawdziwej klasie pozytywnej, natomiast kolor pomarańczowy dla negatywnej. Źródło: Opracowanie własne.

Zauważmy, że maksymalizowana funkcja wiarygodności jest sumą wartości funkcji $J_i(\theta)$ dla poszczególnych obserwacji. W celu lepszego zrozumienia całości przyjrzymy się bliżej własnościom funkcji $J_i(\theta)$.

Wykres 1.2 ze strony 11 przedstawia przyjmowane wartości funkcji $J_i(\theta)$ w zależności od estymowanego prawdopodobieństwa $h_\theta(\boldsymbol{x_i})$. Kolorem niebieskim jest wyznaczony przebieg funkcji dla klasy negatywnej, a pomarańczowym dla pozytywnej. Możemy zauważyć, że funkcja jest symetryczna względem wartości prawdopodobieństwa 0.5 dla obu klas. Oznacza to, że ten klasyfikator przykłada taką samą wagę dla obu rodzajów błędów. Ponadto rozważając następujące granice dla $y_i = 0$

$$\lim_{h_\theta(\boldsymbol{x_i}) \to 0} -\log(1 - h_\theta(\boldsymbol{x_i})) = 0,$$

$$\lim_{h_\theta(\boldsymbol{x_i})\to 1} -\log(1 - h_\theta(\boldsymbol{x_i})) = \infty$$

oraz dla $y_i = 1$

$$\lim_{h_\theta(\boldsymbol{x_i})\to 0} -\log(h_\theta(\boldsymbol{x_i})) = \infty,$$

$$\lim_{h_\theta(\boldsymbol{x_i})\to 1} -\log(h_\theta(\boldsymbol{x_i})) = 0$$

otrzymujemy, że koszt klasyfikacji dla pojedynczej obserwacji przyjmuje w przybliżeniu następujące wartości:

$$J_i(\theta) \approx \begin{cases} 0, & \text{jeżeli } h_\theta(\boldsymbol{x_i}) \approx y_i, \\ \infty, & \text{jeżeli } 1 - h_\theta(\boldsymbol{x_i}) \approx y_i. \end{cases}$$

To znaczy, że w przypadku prawidłowego zaklasyfikowania danej obserwacji funkcja przyjmuje wartość bliską zeru, natomiast w przypadku pomyłki nieskończoności. Stąd możemy wywnioskować, że koszty popełnienia bądź niepopełnienia błędu są następujące:

$$C_{1,1}^{(i)} = C_{0,0}^{(i)} \approx 0,$$

$$C_{1,0}^{(i)} = C_{0,1}^{(i)} \approx \infty.$$

Jest to wynik, który nie oddaje charakteru nierówności kosztu popełniania różnych błędów, stąd powstaje motywacja, aby zmodyfikować podaną funkcję i stworzyć regresję logistyczną, która będzie w stanie modyfikować koszty poszczególnych klasyfikacji.

1.2.2 Drzewo decyzyjne

Drzewo decyzyjne to nieparametryczny model uczenia nadzorowanego, który jest wykorzystywany do regresji oraz klasyfikacji. Jego celem jest nauczenie się na podstawie dostępnych danych prostych reguł decyzyjnych. Do jego największych zalet należy prostota interpretacji otrzymanych wyników oraz mała ilość potrzebnych przygotowań danych.

Wyróżniamy trzy rodzaje algorytmów, które są odpowiedzialne za proces uczenia drzew. Poniżej krótki opis każdego z nich.

- ID3 (Iterative Dichotomiser 3) – algorytm pozwala na tworzenie drzew decyzyjnych, które w poszczególnych węzłach mogą mieć więcej niż dwa podziały. Działa jedynie dla zmiennych kategorycznych. Drzewo rośnie do maksymalnego rozmiaru, a następnie jest przycinane;

- C4.5 – następca ID3, który znosi restrykcję używania tylko kategorycznych atrybutów poprzez dyskretyzację zmiennych ciągłych. Następnie nauczone drzewo zamieniane jest w zestaw reguł, które są szeregowane względem poprawy skuteczności. Przycinanie w tym wypadku polega na usuwanie reguł, które nie poprawiają wyniku modelu;

- CART (z angielskiego *Classification and Regression Tree*) – bardzo podobny algorytm do C4.5, który umożliwia tworzenie drzew regresyjnych (w poprzednich algorytmach mamy do czynienia jedynie z zadaniami klasyfikacyjnymi) oraz nie tworzy zestawu reguł. Konstruuje on binarne drzewa, które do podziału używa atrybutu oraz progu odcięcia, który maksymalizuje przyrost informacji.

Warto zaznaczyć, że implementacja w bibliotece *sklearn* w Pythonie wykorzystuje zmodyfikowaną wersję algorytmu CART, który nie obsługuje zmiennych kategorycznych. Poniżej prezentujemy teorię matematyczną stojącą za implementacją tego algorytmu [9].

Niech $\mathcal{Q}$ będzie zbiorem obserwacji dla danego węzła m. Ponadto niech podział będzie reprezentowany przez parę $\theta = (j, t_m)$, gdzie j to j-ty atrybut wektora $\boldsymbol{x}_i$, a t_m to warunek podziału danych na podzbiory $\mathcal{Q}_l$ oraz $\mathcal{Q}_r$, gdzie

$$\mathcal{Q}_l(\theta) = \{\boldsymbol{x}_i : \boldsymbol{x}_i \in \mathcal{Q} \wedge x_i^j \leq t_m\},$$

$$\mathcal{Q}_r(\theta) = \mathcal{Q} \setminus \mathcal{Q}_l(\theta).$$

Czystość danego węzła jest wyznaczana w następujący sposób

$$G(\mathcal{Q}, \theta) = \frac{|\mathcal{Q}^l|}{|\mathcal{Q}|} I(\mathcal{Q}_l(\theta)) + \frac{|\mathcal{Q}^r|}{|\mathcal{Q}|} I(\mathcal{Q}_r(\theta)),$$

gdzie $I(\cdot)$ może być dowolną funkcją miary zanieczyszczenia. Ostatecznie w danym węźle wybierany jest ten podział, który minimalizuje funkcję czystości

$$\theta^* = \arg\min_{\theta} G(\mathcal{Q}, \theta).$$

Dalej następuje podział na podzbiory $\mathcal{Q}_l(\theta^*)$ i $\mathcal{Q}_r(\theta^*)$ i algorytm działa rekurencyjnie, aż do osiągnięcia maksymalnej głębokości. Ostatnie podziały, które pozostały na końcu drzewa, nazywamy liśćmi i w momencie predykcji klasyfikują one daną próbkę lub zwracają odpowiednie prawdopodobieństwo. W zależności od implementacji oraz preferencji użytkownika po zakończeniu działania algorytmu wykorzystuje się przycinanie drzew. Jest to metoda, która w ogólności polega na usuwanie nadmiarowych węzłów, które mogą powodować zbyt dokładne dopasowanie do danych.

W zależności od rodzaju problemu wykorzystujemy różne miary zanieczyszczenia [9]. Dla zadań klasyfikacyjnych są to

- Misclassification: $I_m(\mathcal{Q}) = 1 - \max(p, 1-p)$,
- Entropy: $I_e(\mathcal{Q}) = -p \log_2 p - (1-p) \log_2 (1-p)$,
- Gini: $I_g(\mathcal{Q}) = 2p(1-p)$,

a dla zadań regresyjnych

- $I_{mse}(\mathcal{Q}) = \dfrac{1}{|\mathcal{Q}|} \sum\limits_{i \in \mathcal{Q}} (y_i - q)^2$,
- $I_{mae}(\mathcal{Q}) = \dfrac{1}{|\mathcal{Q}|} \sum\limits_{i \in \mathcal{Q}} |y_i - q|$,

gdzie:

$$p = \frac{|\{\boldsymbol{x}_i : \boldsymbol{x}_i \in \mathcal{Q} \wedge y_i = 1\}|}{|\mathcal{Q}|},$$

$$q = \frac{\sum_{i \in \mathcal{Q}} y_i}{|\mathcal{Q}|}.$$

1.2.3 Las losowy

Las losowy oraz algorytm XGBoost są przedstawicielami szerokiej klasy modeli typu *ensemble*. Głównym celem tej metodologii jest wprowadzenie losowych perturbacji do zbioru treningowego w celu utworzenia różnych klasyfikatorów bazowych, których wspólne predykcje będą lepsze niż każdego z modeli bazowych osobno. Istnieją ponadto trzy główne powody, dlaczego te metody działają znacznie lepiej niż pojedyncze modele.

- Statystyczny – w przypadku, gdy mamy do czynienia z małym zbiorem danych, istnieje możliwość stworzenia kilku bardzo dobryh klasyfikatorów, które na zbiorze treningowym mają taką samą skuteczność. W przypadku dodatkowego zbioru testowego istnieje szansa wybrania nieoptymalnego;

- Obliczeniowy – większość algorytmów polega na optymalizacji numerycznej, która może zatrzymać się w minimum lokalnym. Dzięki zastosowaniu kilku modeli uzyskujemy szansę przeszukania większej części przestrzeni parametrów modeli;

- Reprezentacyjny – często funkcja f reprezentująca prawdziwy model jest bardzo skomplikowana i niemożliwe jest przedstawienie jej za pomocą jednego modelu. Wykorzystując złożenie kilku modeli bazowych, mamy szansę lepiej przybliżać prawdziwą funkcję f.

W celu utworzenia T klasyfikatorów bazowych należy ze zbioru treningowego $\mathcal{S}$ wybrać S_j dla $j = 1, 2, \ldots, T$ losowych podzbiorów obserwacji, które wykorzystamy do uczenia. W tym celu wykorzystuje się następujące procedury losowania.

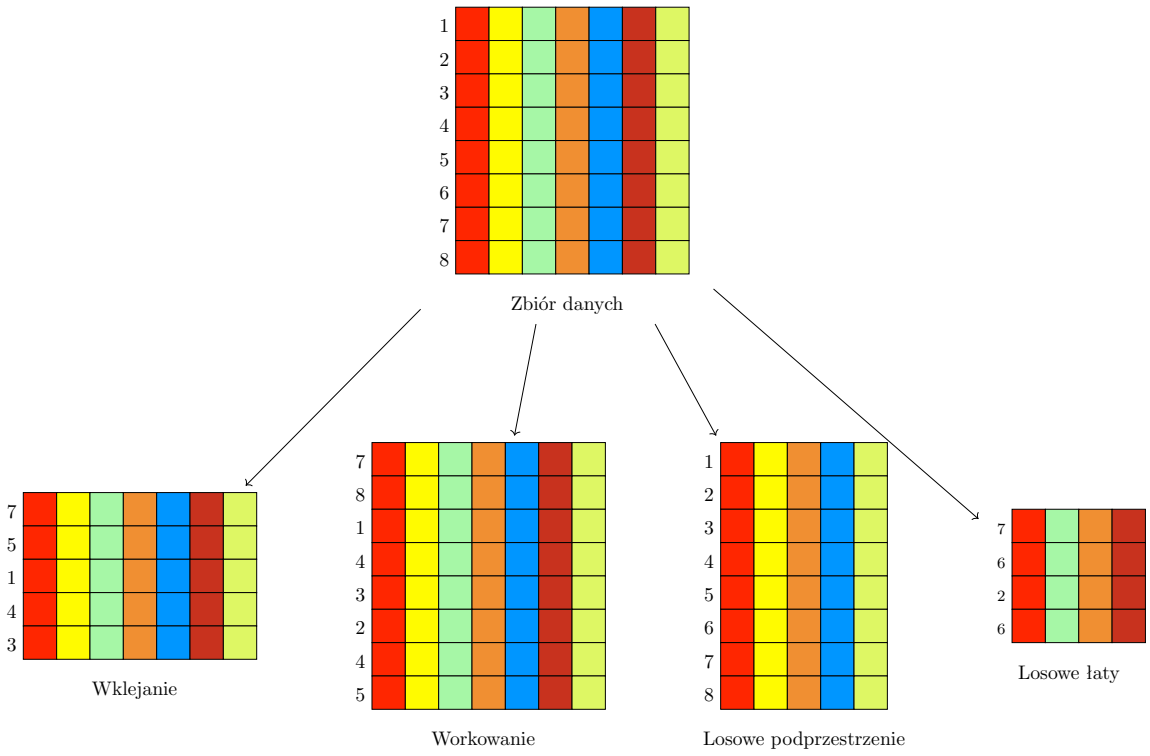
- Wklejanie (*Pasting*) – Polega na niezależnym losowaniu $K < N$ próbek ze zbioru $\mathcal{S}$ bez powtórzeń. Najczęściej wykorzystywana jest na bardzo dużych zbiorów danych, gdzie zasoby obliczeniowe są ograniczone i nie jest możliwe wczytanie wszystkich danych do pamięci komputera. Ponadto dzięki możliwości wykorzystania obliczeń równoległych skraca się czas tworzenia całego modelu. Próbki mogą być losowane z rozkładu jednostajnego bądź wykorzystując metody biorące pod uwagę ważność poszczególnych obserwacji [6].

- Workowanie (*ang. Bagging*) – Polega na niezależnym losowaniu $K \leq N$ próbek ze zbioru $\mathcal{S}$ z powtórzeniami. Najczęściej jest wykorzystywany w przypadku małych zbiorów danych, kiedy chcemy wykorzystać maksymalnie dużą ilość obserwacji, lecz także wprowadzić losowość do kolejnych klasyfikatorów bazowych.

- Losowe podprzestrzenie (*ang. Random Subspaces*) – Metoda polegająca na wybraniu do każdego klasyfikatora bazowego wszystkich obserwacji, lecz losuje się tylko część atrybutów opisujące dane. Głównym celem tego zabiegu jest poprawa skuteczności poprzez zwiększenie różnorodności klasyfikatorów bazowych. [17]

- Losowe łaty (*ang. Random Patches*) – Procedura łącząca metodę wklejania z losowymi podprzestrzeniami. Jej celem jest połączenie zalet obu metod i ostatecznie poprawienie skuteczności modelu z jednoczesną optymalizacją wykorzystywanego czasu oraz pamięci. [13]

Na rysunku 1.3 schematycznie przedstawiono wyżej omówione zasady działania poszczególnych procedur losowania. Należy zwrócić uwagę zarówno na powtarzające się lub nie



Rysunek 1.3: Schemat działania procedur losowania. Źródło: Opracowanie własne.

numery wierszy, jak i na kolory kolumny symbolizujące odpowiednie zmienne z oryginalnego zbioru danych.

Warto zauważyć, że mimo tego, że powyższe metody były oryginalnie tworzone z myślą o wykorzystaniu drzewa decyzyjnego jako klasyfikatora bazowego, to nic nie stoi na przeszkodzie wykorzystać inny model. Ten fakt został między innymi wykorzystany w pracy A.Bahnsena [2], gdzie wykorzystując omawiane w sekcji 1.4.2 drzewa decyzyjne wrażliwe na koszt, stworzono lasy losowe wrażliwe na koszt.

Procedura tworzenia lasu losowego jest podzielona na dwie fazy [7]. Pierwsza z nich wykorzystuje workowanie do wygenerowania odpowiednich próbek dla klasyfikatorów bazowych. Następnie trenujemy zmodyfikowane drzewa decyzyjne, które na etapie tworzenia każdego kolejnego węzła w drzewie losują podzbiór zmiennych do wyboru podziału. Finalnie podczas predykcji każdy model dokonuje klasyfikacji i ostatecznym wynikiem modelu jest większość głosów. Nie jest to jedyna możliwość agregowania wyników, przykładowo wykorzystywana w bibliotece *sklearn* w Python implementacja uśrednia wyniki prawdopodobieństwa z każdego klasyfikatora [9].

1.2.4 XGBoost

W ogólności proces trenowania modeli w uczeniu maszynowym polega na znalezieniu najlepszych parametrów modelu θ, które najlepiej pasują do danych treningowych ze zbioru $\mathcal{S}$ i ich prawdziwych wartości $\boldsymbol{y}$. W celu dokonania tego procesu definiujemy funkcję celu (*ang. objective*), którą w ogólnej postaci możemy zapisać jako

$$\text{Obj}(\theta) = J(\theta) + \Omega(\theta), \tag{1.5}$$

gdzie $J(\theta)$ to funkcja straty, a $\Omega(\theta)$ to wyraz odpowiadający za regularyzację, która nakłada odpowiednie kary na współczynniki modelu, aby zapobiec zbytniemu dopasowaniu do danych. Za funkcję straty najczęściej przyjmujemy w przypadku zadań regresyjnych błąd średniokwadratowy

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2,$$

gdzie $\hat{y}_i$ to predykcja modelu dla i-tej obserwacji, lub w przypadku zadań klasyfikacyjnych entropię krzyżową daną równaniem 1.4.

Algorytm XGBoost (lub z angielskiego *Extreme Gradient Boosting*) to model, który w iteracyjny sposób tworzy kolejne drzewa decyzyjne typu CART (patrz sekcja 1.2.2), które na liściu nie zawierają samej decyzji binarnej, lecz przypisują liczbę rzeczywistą, która daje znacznie bogatszą reprezentację, która wychodzi poza zagadnienia klasyfikacyjne [10]. Podobnie jak w przypadku lasu losowego wykorzystujemy wiele klasyfikatorów bazowych, a następnie dzięki wspomnianej reprezentacji możemy zapisać predykcję i-tej obserwacji jako

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), \quad f_k \in \mathcal{F},$$

gdzie K oznacza liczbę drzew decyzyjnych, f_k funkcje z przestrzeni $\mathcal{F}$ wszystkich możliwych drzew CART. Warto zaznaczyć, że w przypadku zagadnienia regresyjnego predykcja y_i jest bezpośrednim wynikiem, natomiast dla problemów klasyfikacyjnych wykorzystuje się funkcję sigmoidalną (podobnie jak w regresji logistycznej) w celu zamiany wartości liczby rzeczywistej na predykcję z przedziału $[0, 1]$.

Korzystając z ogólnej postaci funkcji celu danej równaniem 1.5 funkcja $\mathrm{Obj}(\theta)$, którą będziemy optymalizować, przyjmuje następującą postać:

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k),$$

gdzie $l(y_i, \hat{y}_i)$ to funkcja straty, przykładowo błąd średniokwadratowy lub entropia krzyżowa, a $\Omega(f_k)$ to regularyzacja nałożona na drzewo f_k.

Uczenie drzew decyzyjnych jest znacznie trudniejszym zadaniem niż standardowe zagadnienia optymalizacyjne, gdyż w tym przypadku nie jesteśmy w stanie wprost policzyć gradientu i wyliczyć kolejnych wartości współczynników. Zamiast tego użyjemy uczenia addytywnego, które polega na naprawianiu błędów z poprzedniej iteracji oraz dodanie kolejnego drzewa. Przez $\hat{y}_i^{(t)}$ będziemy oznaczać predykcję dla i-tej obserwacji w kroku t. Zatem możemy zapisać kolejne predykcje w następujący sposób.

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$$

$$\ldots$$

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

Podczas każdego z kroków iteracji będziemy chcieli minimalizować odpowiednią funkcję celu, która w tym przypadku ma postać

$$\mathrm{Obj}^{(t)}(\theta) = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k).$$

Wykorzystując powyższe wzory, otrzymujemy następujące wyrażenie

$$\text{Obj}^{(t)}(\theta) = \sum_{i=1}^{N} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{const.} \ .$$

Korzystając z rozwinięcia szeregu Taylora dla funkcji straty, otrzymujemy ogólny wzór:

$$\text{Obj}^{(t)}(\theta) = \sum_{i=1}^{N} \left[l(y_i, \hat{y_i}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + \text{const.} \ ,$$

gdzie

$$g_i = \partial_{\hat{y_i}^{(t-1)}} l(y_i, \hat{y_i}^{(t-1)}),$$
$$h_i = \partial^2_{\hat{y_i}^{(t-1)}} l(y_i, \hat{y_i}^{(t-1)}).$$

Następnie po redukcji stałych, które są nieistotne z punktu widzenia optymalizacji, otrzymujemy

$$\sum_{i=1}^{N} [g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)] + \Omega(f_t).$$

Zauważmy, że nigdzie bezpośrednio nie korzystaliśmy z postaci funkcji straty, zatem w ten sposób jesteśmy w stanie zbudować model optymalizujący dowolną dwukrotnie różniczkowalną funkcję ciągłą. Taka uniwersalność powoduje, że algorytm ten znajduje zastosowanie w wielu zadniach z zakresu uczenia maszynowego, jak i innych problemów optymalizacyjnych.

1.3 Klasyfikacja wrażliwa na koszt

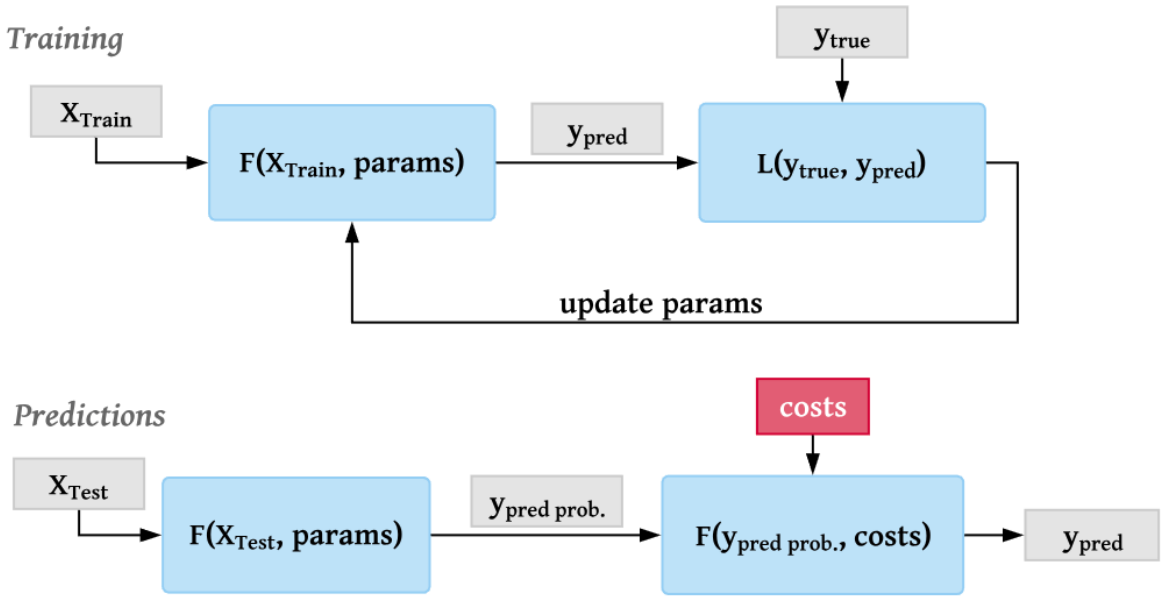
Metody klasyfikacji wrażliwej na koszt (*ang. Cost Dependent Classification*) są przykładem pierwszego rodzaju modeli wrażliwych na koszt. W przypadku tych metod trening zawierający informację o koszcie odbywa się dopiero po etapie stworzenia modelu zwracającego prawdopodobieństwa. Cały proces jest przedstawiony na Rysunku 1.4 na stronie 18. Górna część diagramu przedstawia standardowy proces uczenia modelu predykcyjnego, natomiast dolna część reprezentuje schemat dokonywania predykcji, w którym najpierw estymujemy prawdopodobieństwa dla zbioru testowego, a następnie wykorzystując te wartości oraz koszt klasyfikacji dokonujemy ostatecznej decyzji, czy dana obserwacja jest pozytywna, czy negatywna. Warto wspomnieć, że w celu dokonania kolejnego treningu modelu potrzebujemy podzbioru danych, który nie był wykorzystywany do uczenia podstawowego modelu. Najczęściej taki zbiór nazywamy walidacyjnym.

1.3.1 Optymalizacja progu

Metoda optymalizacji progu (TO z angielskiego *threshold optimization*) jest popularną metodą wyznaczania odpowiedniego progu prawdopodobieństwa, powyżej którego wszystkie obserwacje oznaczamy jako pozytywnie zaklasyfikowane, a poniżej negatywnie. Może być ona wykorzystywana nie tylko do problemów wrażliwych na koszt, lecz do dowolnie zdefiniowanego zagadnienia, w którym potrzebne są zero-jedynkowe predykcje. Jej sformułowanie wygląda następująco

$$\arg\min_{t \in [0,1]} f(\boldsymbol{y}, \boldsymbol{c}),$$

gdzie



Rysunek 1.4: Schemat przedstawiający proces uczenia klasyfikatora wrażliwego na koszt. Źródło: `https://towardsdatascience.com/fraud-detection-with-cost-sensitive-machine-learning-24b8760d35d9`

- $f(\cdot,\cdot)$ - funkcja miary skuteczności modelu, która jako argumenty przyjmuje wektor prawdziwych wartości klasyfikacji oraz wektor binarnych predykcji, np. dokładność,
- $\boldsymbol{c} = (p_i > t)_{i=1}^{n}$ - wektor binarnych klasyfikacji modelu,
- t - wartość progu decyzyjnego.

W naszym przypadku funkcją f będzie funkcja oszczędności (1.2). Innymi słowy, będziemy szukać takiego progu decyzyjnego, który zmaksymalizuje wartość oszczędności.

1.3.2 Minimalizacja ryzyka bayesowskiego

Minimalizacja ryzyka bayesowskiego (BMR z angielskiego *Bayesian Minimum Risk*) to model decyzyjny, którego celem jest zmierzenie oraz porównanie wartości prawdopodobieństw wystąpienia pewnych zdarzeń i kosztów związanych z podjęciem określonych decyzji [5]. Polega na przypisaniu odpowiedniej wartości reprezentującej ryzyko zaklasyfikowania danej obserwacji jako pozytywnej lub negatywnej, które definiujemy w następujący sposób

$$R(c_i = 1|\boldsymbol{x}_i) = L(c_i = 1|y_i = 1)P(y_i = 1|\boldsymbol{x}_i) + L(c_i = 1|y_i = 0)P(y_i = 0|\boldsymbol{x}_i),$$

$$R(c_i = 0|\boldsymbol{x}_i) = L(c_i = 0|y_i = 0)P(y_i = 0|\boldsymbol{x}_i) + L(c_i = 0|y_i = 1)P(y_i = 1|\boldsymbol{x}_i),$$

gdzie

- $P(y_i = 1|\boldsymbol{x}_i)$, $P(y_i = 0|\boldsymbol{x}_i)$ - oznaczają estymowane przez model prawdopodobieństwa zaklasyfikowania obserwacji jako odpowiednio pozytywna oraz negatywna klasa i $P(y_i = 0|\boldsymbol{x}_i) = 1 - P(y_i = 1|\boldsymbol{x}_i)$,
- $L(c_i = j|y_i = k)$ oraz $j, k \in \{0, 1\}$ - funkcja kosztu, która określa stratę, którą poniesiemy w zależności od wyniku poprawności klasyfikacji.

Decyzja dotycząca danej obserwacji jest opisana następującą nierównością

$$R(c_i = 1|\boldsymbol{x}_i) \leq R(c_i = 0|\boldsymbol{x}_i)$$

i oznacza ona, że klasyfikujemy dany przykład jako pozytywny, jeżeli ryzyko takiej decyzji jest mniejsze niż zaklasyfikowania danej obserwacji jako negatywną. Po przeprowadzaniu odpowiednich przekształceń algebraicznych (patrz dodatek A) oraz zakładając, że

$$L(c_i = 0|y_i = 1) - L(c_i = 1|y_i = 1) - L(c_i = 0|y_i = 0) + L(c_i = 1|y_i = 0) > 0$$

otrzymujemy następujący wzór na klasyfikację przykładu jako pozytywny:

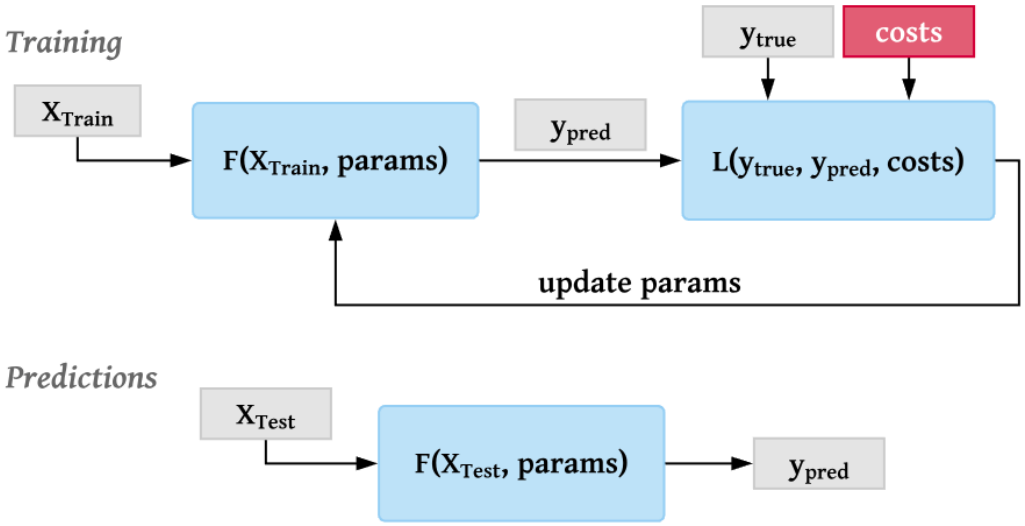
$$P(c_i = 1|\boldsymbol{x}_i) \geq \frac{L(c_i = 1|y_i = 0) - L(c_i = 0|y_i = 0)}{L(c_i = 0|y_i = 1) - L(c_i = 1|y_i = 1) - L(c_i = 0|y_i = 0) + L(c_i = 1|y_i = 0)}.$$

W przypadku gdy za funkcję kosztu przyjmiemy odpowiednie wartości z macierzy kosztu, to otrzymamy następujący próg decyzyjny

$$P(c_i = 1|\boldsymbol{x}_i) \geq \frac{C^{(i)}_{1,0} - C^{(i)}_{0,0}}{C^{(i)}_{0,1} - C^{(i)}_{1,1} - C^{(i)}_{0,0} + C^{(i)}_{1,0}}.$$

1.4 Trening wrażliwy na koszt

Drugą podgrupą metod wrażliwych na koszt jest trening wrażliwy na koszt (*ang. Cost Sensitive Trainig*). Są to metody, które już na etapie treningu modelu biorą pod uwagę koszt związany z klasyfikacją danej obserwacji. Schemat uczenia modelu jest przedstawiony na wykresie 1.5 na stronie 19. Model jako wejście otrzymuje zbiór danych treningowych, następnie dokonuje predykcji i na bazie prawdziwych odpowiedzi oraz kosztów wyznaczana jest skuteczność modelu. W następnej kolejności aktualizowane są wagi, a cały proces jest iteracyjnie powtarzany aż do momentu osiągnięcia zadanego kryterium stopu.



Rysunek 1.5: Schemat przedstawiający proces uczenia modelu wrażliwego na koszt. Źródło: `https://towardsdatascience.com/fraud-detection-with-cost-sensitive-machine-learning-24b8760d35d9`

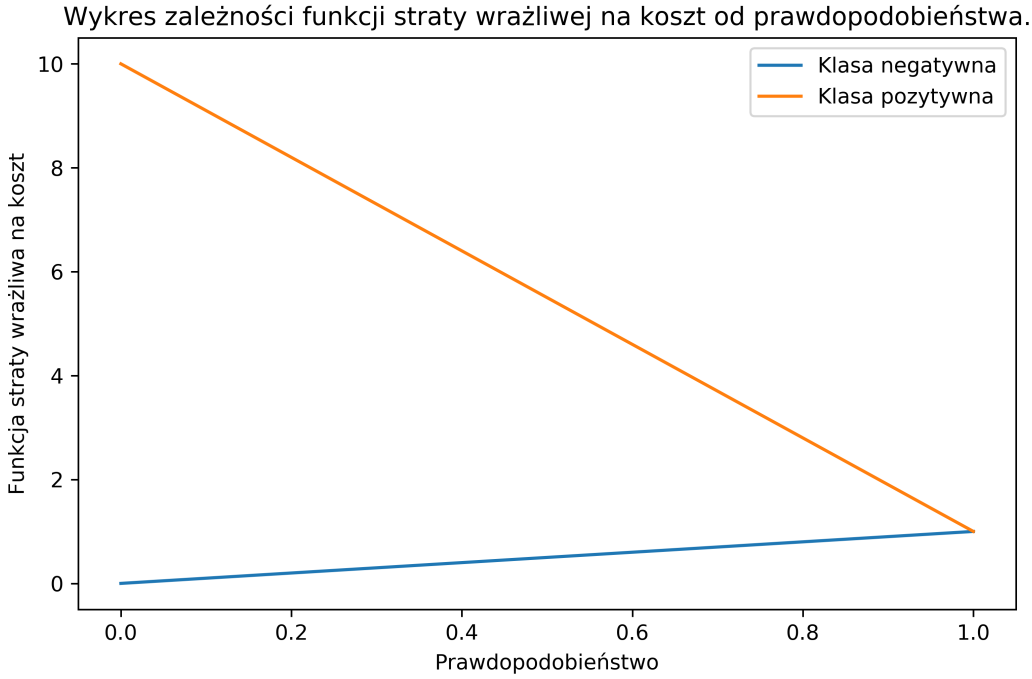
1.4.1 Regresja logistyczna wrażliwa na koszt

Pierwszy modelem, który stosunkowo łatwo przystosować do bycia wrażliwym na koszt jest regresja logistyczna. Kontynuując rozważania z podrozdziału 1.2.1, chcemy, aby funkcja straty przyjmowała następujące wartości, które odpowiadają wartościom z macierzy kosztu:

$$J_i^c(\theta) = \begin{cases} C_{1,1}^{(i)}, & \text{jeżeli } y_i = 1 \text{ i } h_\theta(\boldsymbol{x_i}) \approx 1, \\ C_{0,0}^{(i)}, & \text{jeżeli } y_i = 0 \text{ i } h_\theta(\boldsymbol{x_i}) \approx 0, \\ C_{1,0}^{(i)}, & \text{jeżeli } y_i = 0 \text{ i } h_\theta(\boldsymbol{x_i}) \approx 1, \\ C_{0,1}^{(i)}, & \text{jeżeli } y_i = 1 \text{ i } h_\theta(\boldsymbol{x_i}) \approx 0. \end{cases}$$

W rezultacie funkcją, która zachowuje się w następujący sposób, jest

$$J_i^c(\theta) = y_i\Big(h_\theta(\boldsymbol{x_i})C_{1,1}^{(i)} + (1 - h_\theta(\boldsymbol{x_i}))C_{0,1}^{(i)}\Big) + (1 - y_i)\Big(h_\theta(\boldsymbol{x_i})C_{1,0}^{(i)} + (1 - h_\theta(\boldsymbol{x_i}))C_{0,0}^{(i)}\Big).$$



Rysunek 1.6: Wykres zależności wrażliwej na koszt funkcji straty od prawdopodobieństwa predykcji danej obserwacji. Kolor niebieski przedstawia wykres dla próbki o prawdziwej klasie pozytywnej, natomiast kolor pomarańczowy dla negatywnej. Wykres dla przykładowych wartości: $C_{\mathrm{TP}} = 1$, $C_{\mathrm{FN}} = 10$, $C_{\mathrm{FP}} = 1$, $C_{\mathrm{TN}} = 0$. Źródło: Opracowanie własne.

Wykres funkcji $J_i^c(\theta)$ jest przedstawiony na wykresie 1.6 na stronie 20. Kolorem niebieskim zaznaczona jest predykcja dla klasy negatywnej, a pomarańczowym dla pozytywnej. Możemy zauważyć pożądaną przez nas asymetrię popełniania błędów oraz dokładne wartości, jakie funkcja przyjmuje podczas odpowiednich rodzajów błędów.

Ostatecznie funkcją, którą będziemy minimalizować, jest

$$J^c(\theta) = \sum_{i=1}^{N} J_i^c(\theta).$$

Następnie podobnie jak w przypadku standardowej regresji logistycznej wykorzystujemy odpowiedni algorytm optymalizujący, który znajdzie odpowiednie współczynniki modelu.

1.4.2 Drzewo decyzyjne wrażliwe na koszt

Analogicznie jak w przypadku regresji logistycznej w celu wprowadzenia kosztu do treningu drzewa decyzyjnego musimy uzależnić proces powstawania modelu od kosztu danej próbki. Naturalnym miejscem dla drzewa decyzyjnego jest moment podziału danego węzła. Do tej pory wszystkie miary zanieczyszczenia skupiały się na możliwe najlepszym rozróżnieniu próbek w kontekście minimalizacji ilości błędów klasyfikacji. Zamiast tego, w przypadku metody wrażliwej na koszt, naszym celem jest minimalizacja kosztu. W tym celu definiujemy następującą miarę zanieczyszczenia

$$I_c(\mathcal{Q}) = \text{Koszt bazowy}(\boldsymbol{y}^{\mathcal{Q}}, \boldsymbol{C}^{\mathcal{Q}}),$$

gdzie górny indeks $\mathcal{Q}$ oznacza odpowiedni podzbiór obserwacji, które wybieramy do wektora $\boldsymbol{y}$ oraz $\boldsymbol{C}$. Intuicja stojąca za takim sformułowaniem jest następująca. Klasyfikujemy wszystkie przypadki w rozważanym podziale jako pozytywne, następnie jako negatywne i sprawdzamy, która z decyzji miała mniejszy sumaryczny koszt. W momencie predykcji ostateczny werdykt, który może być tylko binarny, zależy od tego, która klasyfikacja minimalizuje sumaryczny koszt na liściu.

Rozdział 2

Eksperyment

W celu porównania klasycznych oraz wrażliwych na koszt metod predykcyjnych przeprowadzimy eksperyment na danych dotyczących detekcji oszustw na kartach płatniczych. Naszym celem będzie sprawdzenie, czy metody wrażliwe na koszt dają lepsze rezultaty niż standardowe modele. Jeżeli tak, to czy dzieje się tak dla każdej z metod, czy tylko dla niektórych. Ponadto będziemy chcieli odpowiedzieć na pytanie, czy jeżeli nastąpiła poprawa wyników względem oszczędności, to czy dzieje się to kosztem skuteczności w sensie standardowych miar.

Całość pracy rozpoczniemy od omówienia zawartości oraz pochodzenia zbioru danych. Następnie przejdziemy do opisu metodologii przeprowadzania eksperymentu. Na końcu omówimy otrzymane wyniki oraz wyciągnięte wnioski.

2.1 Dane

Do eksperymentu zostanie wykorzystany zbiór danych Credit Card Fraud Detection[1], który był stworzony w celu badań naukowych grupy Worldline and the Machine Learning Group[2] z Université Libre de Bruxelles. Zawiera on 284,807 transakcji w tym zaledwie 492 oszustwa. Tabela składa się z 30 kolumn, w tym 28 z nich to zanonimizowane zmienne numeryczne, które były wcześniej poddane transformacji PCA (*ang. Principal Component Analysis*), a pozostałe dwie kolumny to informacje dotyczące godziny oraz kwoty transakcji. Ponadto wśród danych nie ma brakujących wartości. Podczas modelowania pominiemy zmienną czasową, ponieważ sama w sobie nie zawiera ona istotnej informacji, a tworzenie znaczących zmiennych nie jest istotą przeprowadzanego eksperymentu.

Mimo tego, że dane zostały poddane anonimizacji, to na podstawie artykułu A. Bahnsena możemy domyślać się, z jakimi zmiennymi mieliśmy do czynienia przed transformacjami [5]. Podczas procesu dokonywania transakcji standardowo zbierane są:

- data dokonania transakcji,
- numer konta,
- numer karty,
- typ transakcji (płatność internetowa, płatność stacjonarna, wypłata z bankomatu),

[1] Źródło: `https://www.kaggle.com/mlg-ulb/creditcardfraud`
[2] `http://mlg.ulb.ac.be`

- kwota,
- ID beneficjenta transakcji,
- grupa beneficjenta transakcji (przykładowo linie lotnicze, hotel, wypożyczalnia samochodów itp.),
- kraj dokonania transakcji,
- kraj zamieszkania właściciela karty,
- typ karty (przykładowo VISA, MasterCard itp.),
- wiek klienta,
- płeć klienta,
- bank klienta,
- historyczna informacja czy dana transakcja była oszustwem.

Na podstawie wymienionych informacji tworzy się agregaty czasowe bazujące na historii, których celem jest zebranie informacji dotyczących nawyków zakupowych danego klienta[3]. Stworzenie takiej zmiennej polega na zgrupowaniu transakcji w ciągu pewnego odcinka czasowego najpierw według id klienta lub karty kredytowej, dalej przez typ transakcji, beneficjenta transakcji, kategorię beneficjenta lub kraj, kończąc agregacją poprzez zsumowanie ilości bądź wartości tych płatności. Przykładowymi zmiennymi, które powstają w tym procesie, są ilość transakcji dla tego samego klienta w ciągu ostatnich 6 godzin, średnia wartość transakcji w sklepach spożywczych dla danej karty kredytowej z ostatniego tygodnia.

Na podstawie wiedzy dotyczącej modelowania detekcji fraudów z artykułu A. Bahnsena [2] w tabeli 2.1 na stronie 24 definiujemy macierz kosztu dla tego eksperymentu. Wartość $C_a = 1$ to stały koszt administracyjny podjęcia sprawy przez analityka, który sprawdza, czy dana obserwacja jest faktycznie oszustwem, niezależnie od tego, jaki był jego ostateczny werdykt, Amt_i to wartość transakcji, jaką utracimy w przypadku niewskazanie fałszywej transakcji jako oszustwa, natomiast zerowa wartość kosztu dla normalnej obserwacji, która była prawidłowo wskazana, pochodzi z braku konieczności podjęcia inwestygacji oraz strat z tego wynikających.

	Stan pozytywny $y_i = 1$	Stan negatywny $y_i = 0$
Predykcja pozytywna $c_i = 1$	$C^{(i)}_{1,1} = \mathrm{C_a} = 1$	$C^{(i)}_{1,0} = \mathrm{C_a} = 1$
Predykcja negatywna $c_i = 0$	$C^{(i)}_{0,1} = \mathrm{Amt_i}$	$C^{(i)}_{0,0} = 0$

Tabela 2.1: Macierz kosztu eksperymentu.

[3]Wystąpienie A.Bahnsena podczas Konferencji Analytics 2013. Źródło: `https://www.youtube.com/watch?v=YCNkxaVDiA0`

2.2 Opis eksperymentu

Inspiracja do eksperymentu pochodzi z pracy A.Bahnsena dotyczącej porównania metod wrażliwych na koszt dla pięciu rzeczywistych zbiorów danych pochodzących z czterech dziedzin zastosowań: detekcja oszustw na kartach kredytowych, prognoza rezygnacji, ryzyko kredytowe oraz marketing bezpośredni [2]. Naszym celem będzie porównanie standardowych metod oraz metod wrażliwych na koszt na innym zbiorze danych. W tym celu wykorzystamy regresję logistyczną, drzewo decyzyjne, las losowy, XGBoost, drzewo decyzyjne wrażliwe na koszt, a także standardowe modele rozszerzymy metodą optymalizacji progu (TO) oraz minimalizacji ryzyka bayesowskiego (BMR). Dodatkową modyfikacją w stosunku do oryginalnego eksperymentu będzie użycie algorytmu XGBoost, który w wielu przypadkach uzyskuje znacznie lepsze wyniki od reszty modeli uczenia maszynowego, a także spowodowanie, aby brał pod uwagę koszty pomyłek poprzez użycie klasyfikacji wrażliwej na koszt.

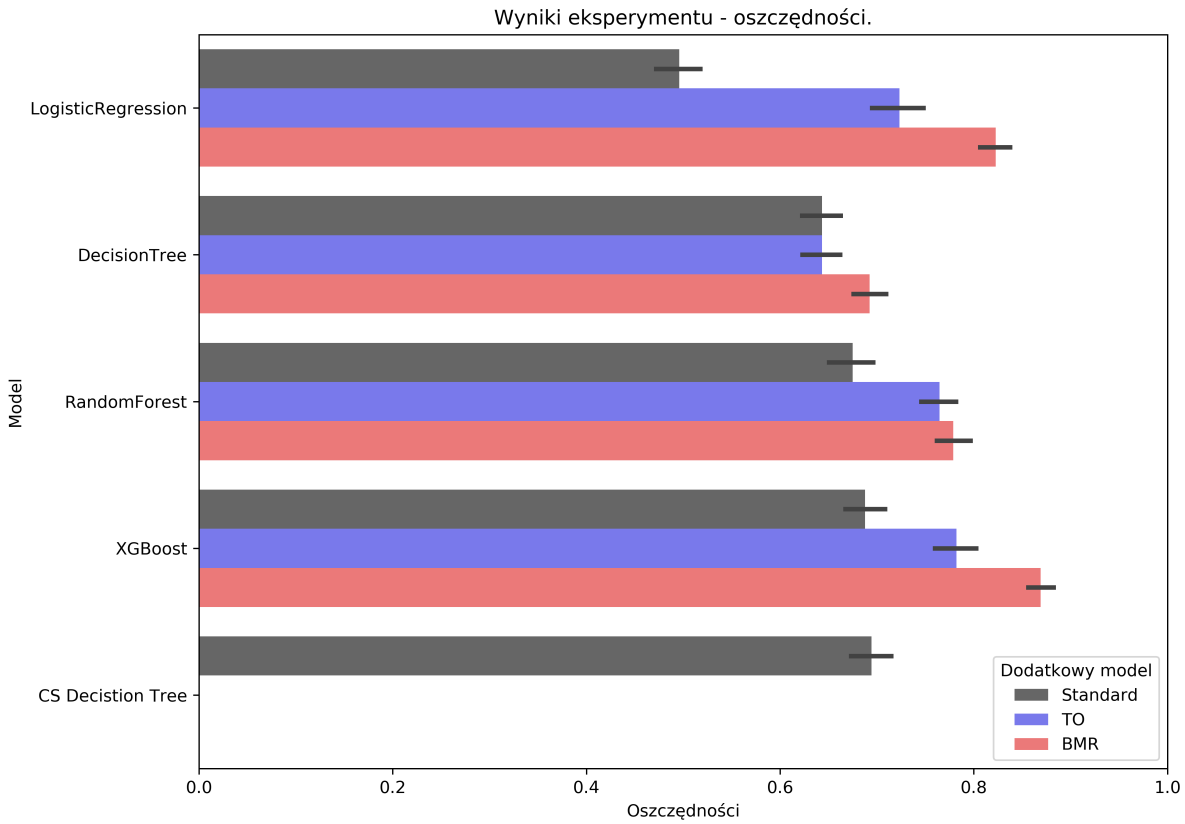
W celu przeprowadzenia eksperymentu pięćdziesięciokrotnie dzielimy zbiór danych w proporcjach 50:17:33 na odpowiednio zbiór treningowy, walidacyjny oraz testowy. Losowanie nowych podziałów zbioru ma na celu zmniejszenie ryzyka, że wynik będzie zależny od wylosowanej próbki. Ponadto wykorzystano tę technikę zamiast standardowej warstwowej walidacji krzyżowej (*ang. Stratified Cross Validation*) z uwagi na chęć uniknięcia zbyt małej próbki testowej, która powstaje w przypadku wykorzystania walidacji krzyżowej z dużą ilością podziałów. Następnie uczymy wszystkie modele standardowe oraz drzewo decyzyjne wrażliwe na koszt na zbiorze treningowym. Dla modelu XGBoost wykorzystujemy zbiór walidacyjny do procesu wczesnego zatrzymywania (*ang. Early stopping*), natomiast dla modeli BMR oraz TO wykorzystujemy ten zbiór jako treningowy. Następnie dla wszystkich modeli dokonujemy predykcji na zbiorze testowym i mierzymy skuteczność typowań, wykorzystując miarę skuteczności oszczędności oraz F_1 Score. Ostatecznie wyniki z wszystkich pięćdziesięciu prób uśredniamy oraz obliczamy wartość odchylenia standardowego.

Do implementacji skryptów został wykorzystany język programowania Python wraz z bibliotekami *costcla*, *sklearn*, *pandas*, *numpy*, *matplotlib* oraz język programowania bash.

2.3 Wyniki

Wyniki eksperymentu dla miary skuteczności oszczędności są zaprezentowane na wykresie 2.1 na stronie 26. Szare, niebieskie oraz czerwone słupki oznaczają dla pierwszych czterych modeli: regresji logistycznej (LogisticRegression), drzewa decyzyjnego (DecisionTree), lasu losowego (RandomForest) oraz XGBoosta, odpowiednio model standardowy, model optymalizacji progu oraz model wykorzystujący minimalizację ryzyka bayesowskiego. Dla modelu drzewa decyzyjnego wrażliwego na koszt (CS Decision Tree) posiadamy jedynie szary słupek, ponieważ model ten sam w sobie jest już wrażliwy na koszt i nie potrzebuje dodatkowych modyfikacji.

Możemy zauważyć, że wśród standardowych modeli oraz drzewa decyzyjnego wrażliwego na koszt najmniejszy wynik oszczędności uzyskała regresja logistyczna a najlepsze las losowy, XGBoost oraz drzewo decyzyjne wrażliwe na koszt. Dodatkowo zwraca uwagę fakt, że wrażliwa na koszt wersja drzewa decyzyjnego przynosi lepsze rezultaty niż standardowy model. W przypadku, gdy weźmiemy pod uwagę rozszerzone wersje modeli standardowych o klasyfikację wrażliwą na koszt, to zdecydowanie najlepszym modelem staje się XGBoost w połączeniu z minimalizacją ryzyka bayesowskiego. Warto zauważyć, że w przypadku optymalizacji progu wartość oszczędności jest co najmniej równa bądź większa niż wynik

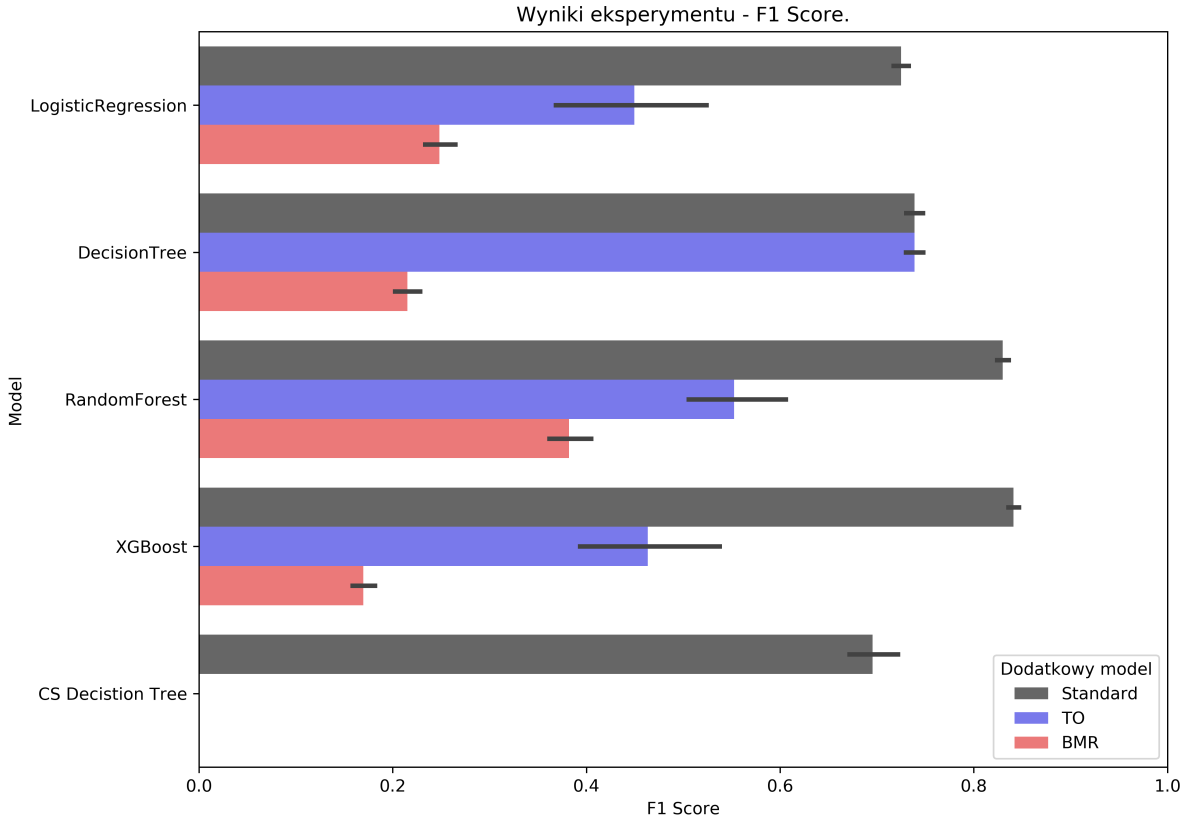


Rysunek 2.1: Wyniki eksperymentu dla miary skuteczności oszczędności. Szare, niebieskie oraz czerwone słupki oznaczają dla pierwszych czterych modeli: regresji logistycznej (LogisticRegression), drzewa decyzyjnego (DecisionTree), lasu losowego (RandomForest) oraz XGBoosta, odpowiednio model standardowy, model optymalizacji progu oraz model wykorzystujący minimalizację ryzyka bayesowskiego. Dla modelu drzewa decyzyjnego wrażliwego na koszt (CS Decision Tree) mamy jedynie jedną wartość. Wysokość słupka reprezentuje wartość średnią z 50 powtórzeń, a długość dołączonej kreski oznacza ± wartość odchylenia standardowego w odpowiednią stronę. Źródło: Opracowanie własne.

uzyskany przez standardowy model. Dla minimalizacji ryzyka bayesowskiego zachodzi podobna sytuacja, gdzie w każdym przypadku wyniki są lepsze. Ponadto średnia wartość oszczędności tego modelu BMR w obrębie każdego z modeli standardowych jest najwyższa. Podsumowując, w przypadku tego eksperymentu, gdyby najważniejsza byłaby dla nas wartość oszczędności, to najlepszym wyborem byłby algorytm XGBoost z dodatkowym modelem minimalizacji ryzyka bayesowskiego. Dodatkowo w każdym z przypadków model wrażliwy na koszt poprawia wynik oszczędności. Spójrzmy jednak, jak zmienia się ogólna moc predykcyjna modelu.

W tym celu przeanalizujemy wyniki eksperymentu dla miary skuteczności F_1 Score, które są zaprezentowane na wykresie 2.2 na stronie 26. Oznaczenia są analogiczne jak do wykresu dotyczącego oszczędności. Możemy zauważyć, że dla standardowych wersji modeli w przypadku tej miary skuteczności najlepsze rezultaty osiągnął las losowy oraz XGBoost. Dodatkowo model drzewa decyzyjnego wrażliwego na koszt uzyskał w tym wypadku gorsze rezultaty niż standradowa wersja. W przypadku klasyfikacji wrażliwej na koszt modele optymalizacji progu prawie zawsze pogarszały wyniki (z wyjątkiem drzewa decyzjnego, gdzie wynik był taki sam jak bez modelu), a dla minimalizacji ryzyka bayesowskiego wynik

zawsze był gorszy od standardowej wersji modelu. Intuicyjnie jest to zrozumiały rezultat, ponieważ nasz model przestał zwracać uwagę na transakcje, które miały małą wartość, nawet jeżeli przypadek ten był łatwy do wykrycia, a profilaktycznie zaczął sprawdzać duże transakcje, które mogły mieć nawet małe prawdopodobieństwa bycia oszustwem. W związku z tym modele zaczęły oszczędzać więcej pieniędzy kosztem ogólnej skuteczności modelu, przez to zwracają większą ilości fałszywie pozytywnych klasyfikacji.



Rysunek 2.2: Wyniki eksperymentu dla miary skuteczności F1-Score. Szare, niebieskie oraz czerwone słupki oznaczają dla pierwszych czterych modeli: regresji logistycznej (LogisticRegression), drzewa decyzyjnego (DecisionTree), lasu losowego (RandomForest) oraz XGBoosta, odpowiednio model standardowy, model optymalizacji progu oraz model wykorzystujący minimalizację ryzyka bayesowskiego. Dla modelu drzewa decyzyjnego wrażliwego na koszt (CS Decision Tree) mamy jedynie jedną wartość. Wysokość słupka reprezentuje wartość średnią z 50 powtórzeń, a długość dołączonej kreski oznacza $\pm$ wartość odchylenia standardowego w odpowiednią stronę. Źródło: Opracowanie własne.

Zestawiając uzyskane rezultaty z obu miar, możemy dojść do wniosku, że jeżeli najważniejszym i jedynym kryterium, które chcemy rozważać w ocenie jakości modelu, są oszczędności, to w przypadku tego zbioru danych najlepszym wyborem jest algorytm XGBoost z minimalizacją ryzyka bayesowskiego. Natomiast jeżeli poza oszczędnościami chcemy wziąć także pod uwagę ogólną jakość predykcyjną modelu, to wyniki nie jest aż tak jednoznaczny. W zależności od przypadku będziemy musieli wybierać pomiędzy ogólną skutecznością a oszczędnościami. Ciekawym wyborem w tym przypadku może być drzewo decyzyjne wrażliwe na koszt, las losowy lub XGBoost w wersji standardowej. Dodatkową informacją, która mogłaby nam pomóc podjąć decyzję, może być ilość pozytywnych klasyfikacji z każdego modelu oraz ilość analityków zajmujących się sprawdzaniem oszustw,

ponieważ może się okazać, że w wyniku zbyt małej ilości osób nie będziemy w stanie sprawdzić wszystkich wytypowanych spraw. W takiej sytuacji model, który typuje mniej transakcji do sprawdzania przy zachowaniu podobnej skuteczności dla obu miar, może okazać się najlepszym wyborem. Z drugiej strony, jeżeli taka informacja byłaby dostępna na początku modelowania, to wymusiłaby ona dodatkowe założenia oraz inne podejście do procesu doboru metryk.

Na końcu warto dodać, że pierwsze eksperymenty były wykonywane również z wykorzystaniem regresji logistycznej wrażliwej na koszt, natomiast z nieznanych nam przyczyn (być może błędy implementacyjne) uzyskiwała ona ujemne wartości oszczędności oraz F_1 Score bliski zeru. Z tego powodu uzyskane wyniki nie zostały opublikowane w podsumowaniu eksperymentu, ponieważ nie wnosiły one żadnej użytecznej informacji w kontekście przeprowadzanego rozumowania. Ponadto poza tokiem przeprowadzonych prac został zrealizowany dodatkowy eksperyment na innym zbiorze danych i model ten uzyskiwał prawidłowe rezultaty, zatem wina może leżeć także po stronie naszego zbioru danych i/lub definicji macierzy kosztu.

Podsumowanie

Pracę rozpoczęliśmy od wprowadzenia problemu detekcji oszustw na kartach kredytowych oraz omówienia aktualnego podejścia do tworzenia systemów wykrywających oszustwa. Następnie omówiliśmy rodzaje miar, które są standardowo wykorzystywane w statystyce oraz uczeniu maszynowym, a także miary, które stosuje się w problemach wrażliwych na koszt. Dalej opisaliśmy podstawy teoretyczne standardowych modeli predykcyjnych, ich modyfikacji wrażliwych na koszt oraz klasyfikatorów wrażliwych na koszt. W ostatniej części pracy przeprowadziliśmy eksperyment, którego celem było porównanie wybranych klasyfikatorów na danych rzeczywistych dotyczących detekcji oszustw na kartach kredytowych.

Na podstawie przeprowadzonego eksperymentu ustaliliśmy, że na tym konkretnym zbiorze danych dzięki wykorzystaniu metod wrażliwych na koszt bylibyśmy w stanie zwiększyć oszczędności w problemie detekcji oszustw na kartach kredytowych. Najlepszym modelem pod tym względem okazał się XGBoost, który do tej pory nie był rozważanych w pracach A. Bahnsena dotyczących podobnych zagadnień. Dodatkowo zauważyliśmy, że wykorzystanie klasyfikatorów wrażliwych na koszt znacząco obniża skuteczność modelu w sensie standardowych miar. W przypadku, gdybyśmy chcieli zachować balans pomiędzy tymi dwoma miarami, najlepszych wyborem byłby algorytm XGBoost, las losowy lub drzewo decyzyjne wrażliwe na koszt.

Na podstawie przeprowadzonych badań wydaje się, że bardzo ciekawym kierunkiem do dalszych rozważań mogą być metody typu *ensemble*, których klasyfikatorem bazowym byłoby drzewo decyzyjne wrażliwe na koszt. Takie działania zostały już podjęte dla lasów losowych w cytowanej wcześniej pracy A. Bahnsena [2]. Z uwagi na fakt, że algorytm XGBoost otrzymywał najlepsze rezultaty dla obu miar skuteczności oraz drzewo decyzyjne w wersji wrażliwej na koszt uzyskiwało lepsze rezultaty niż standradowa wersja modelu, daje to nadzieje, że XGBoost wrażliwy na koszt mógłby dawać jeszcze lepsze wyniki.

Dodatek A

Minimalizacja ryzyka bayesowskiego

W tej części dodatku wyprowadzimy wzór

$$P(c_i = 1|\boldsymbol{x}_i) \geq \frac{L(c_i = 1|y_i = 0) - L(c_i = 0|y_i = 0)}{L(c_i = 0|y_i = 1) - L(c_i = 1|y_i = 1) - L(c_i = 0|y_i = 0) + L(c_i = 1|y_i = 0)}.$$

Przypomnijmy, że klasyfikujemy daną obserwację jako pozytywną, jeżeli prawdziwa jest nierówność

$$R(c_i = 1|\boldsymbol{x}_i) \leq R(c_i = 0|\boldsymbol{x}_i). \tag{A.1}$$

Ponadto w sekcji 1.3.2 definiowaliśmy ryzyko klasyfikacji w następujący sposób

$$R(c_i = 1|\boldsymbol{x}_i) = L(c_i = 1|y_i = 1)P(y_i = 1|\boldsymbol{x}_i) + L(c_i = 1|y_i = 0)P(y_i = 0|\boldsymbol{x}_i),$$

$$R(c_i = 0|\boldsymbol{x}_i) = L(c_i = 0|y_i = 0)P(y_i = 0|\boldsymbol{x}_i) + L(c_i = 0|y_i = 1)P(y_i = 1|\boldsymbol{x}_i).$$

W celu skrócenia zapisu wprowadźmy jeszcze następujące oznaczenia:

$$P_k = P(y_i = k|\boldsymbol{x}_i), \quad k \in \{0, 1\}$$

$$L_{m,n} = L(c_i = m|y_i = n), \quad m, n \in \{0, 1\}$$

Korzystając z powyższych oznaczeń, definicji $R(c_i = 1|\boldsymbol{x}_i), R(c_i = 0|\boldsymbol{x}_i)$ oraz równania A.1 otrzymujemy

$$L_{1,1}P_1 + L_{1,0}P_0 \leq L_{0,0}P_0 + L_{0,1}P_1.$$

Korzystając z własności

$$P_0 = 1 - P_1$$

mamy

$$L_{1,1}P_1 + L_{1,0}(1 - P_1) \leq L_{0,0}(1 - P_1) + L_{0,1}P_1.$$

Następnie dokonując przekształcenia algebraiczne, dostajemy

$$L_{1,0} - L_{0,0} \leq P_1(L_{0,1} - L_{0,0} - L_{1,1} + L_{1,0}).$$

Zakładając, że $L_{0,1} - L_{0,0} - L_{1,1} + L_{1,0} > 0$ ostatecznie otrzymujemy

$$P_1 \geq \frac{L_{1,0} - L_{0,0}}{L_{0,1} - L_{0,0} - L_{1,1} + L_{1,0}}. \quad \square$$

Bibliografia

[1] Bahnsen, A. C., Aouada, D., Ottersten, B. Example-dependent cost-sensitive logistic regression for credit scoring.

[2] Bahnsen, A. C., Aouada, D., Ottersten, B. Ensemble of example-dependent cost-sensitive decision trees, 2015.

[3] Bahnsen, A. C., Aouada, D., Ottersten, B. Example-dependent cost-sensitive decision trees. *Expert Syst. Appl. 42*, 19 (Nov. 2015), 6609–6619.

[4] Bahnsen, A. C., Stojanovic, A., Aouada, D., Ottersten, B. *Improving Credit Card Fraud Detection with Calibrated Probabilities*. pp. 677–685.

[5] Bahnsen, A. C., Stojanovic, A., Aouada, D., Ottersten, B. Cost sensitive credit card fraud detection using bayes minimum risk. In *2013 12th International Conference on Machine Learning and Applications* (Dec 2013), vol. 1, pp. 333–338.

[6] Breiman, L. Pasting small votes for classification in large databases and on-line. *Machine Learning 36*, 1 (Jul 1999), 85–103.

[7] Breiman, L. Random forests. *Machine Learning 45*, 1 (Oct 2001), 5–32.

[8] Brodersen, K. H., Ong, C. S., Stephan, K. E., Buhmann, J. M. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition* (Aug 2010), pp. 3121–3124.

[9] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* (2013), pp. 108–122.

[10] Chen, T., Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, ACM, pp. 785–794.

[11] Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems* (Berlin, Heidelberg, 2000), Springer Berlin Heidelberg, pp. 1–15.

[12] Elkan, C. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2* (San Francisco, CA, USA, 2001), IJCAI'01, Morgan Kaufmann Publishers Inc., pp. 973–978.

[13] Louppe, G., Geurts, P. Ensembles on random patches. pp. 346–361.

[14] Moser, R. Fraud detection with cost-sensitive machine learning, 2019.

[15] Powers, D., Ailab. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *J. Mach. Learn. Technol 2* (01 2011), 2229–3981.

[16] Sheng, V. S., Ling, C. X. Thresholding for making classifiers cost-sensitive. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1* (2006), AAAI'06, AAAI Press, pp. 476–481.

[17] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 8 (Aug 1998), 832–844.