INTERPRETOWALNOŚĆ I WYJAŚNIALNOŚĆ UCZENIA MASZYNOWEGO

Dr Robert Małysz

WYKŁAD 7 - AGENDA

Metody przykładowe, oparte na regułach oraz kohortowe

1. Counterfactuals (Kontrfaktyczne przykłady)
2. Anchors (Reguły kotwiczące)

WYKŁAD 7 – TAKSNOMIA XAI

WYMIAR 1: SCOPE - ZAKRES WYJAŚNIENIA

Global Methods

Cel: Zrozumienie całego modelu

- PDP / ICE / ALE
- Permutation Importance
- H-statistic
- Global Surrogate
- SHAP Summary

Local Methods

Cel: Wyjaśnienie pojedynczej predykcji

- LIME
- SHAP (local)
- Counterfactuals
- Anchors
- Prototypes

Cohort Methods

Cel: Analiza grup obserwacji

- ICE clusters
- Subgroup analysis
- Group SHAP
- Segment-wise PDP

WYKŁAD 7 – TAKSNOMIA XAI

WYMIAR 3: TECHNIQUE – TECHNIKA WYJAŚNIANIA

Feature Attribution

SHAP - Shapley values

LIME - Local surrogate

Integrated Gradients

LRP - Layer-wise Relevance

Partial Dependence

PDP - Partial Dependence

ICE - Individual curves

ALE - Accumulated effects

SHAPDP - SHAP dependence

Feature Importance

Permutation - Shuffle & measure

Gini/Entropy - Tree-based

Drop-column - Retrain

SHAP - Mean |SHAP|

Example-based

Counterfactuals - DiCE, WhatIf

Prototypes - Representative

Influential - Training impact

Adversarial - Edge cases

Concept-based

TCAV - Concept vectors

CBM - Concept bottleneck

ACE - Automatic concepts

NetDissect - Unit semantics

Surrogate Models

Global - Entire model

Local (LIME) - Per instance

Anchors - Rule-based

Distillation - Teacher-student

WYKŁAD 7 – COUNTERFACTUALS WPROWADZENIE

Definicja

Counterfactual (przykład kontrfaktyczny) to najbliższa możliwa modyfikacja cechy wejściowej instancji, która zmienia predykcję modelu na pożądaną klasę.

Intuicja

"Co musiałoby się zmienić w danych wejściowych, aby otrzymać inną predykcję?"

Przykład z aplikacją kredytową:

- **Obecna sytuacja:** Kredyt odrzucony
- **Pytanie:** Co muszę zmienić, aby otrzymać kredyt?
- **Counterfactual:** "Gdyby Twój roczny dochód był o 5000 PLN wyższy ORAZ gdybyś miał o 2 lata dłuższą historię kredytową, Twój kredyt zostałby zatwierdzony"

WYKŁAD 7 – COUNTERFACTUALS PROBLEM OPTYMALIZACYJNY

Problem optymalizacyjny

$$x' = \arg\min_{x'} \left[L(f(x'), y_{target}) + \lambda \cdot d(x, x') + \sum_{j=1}^{p} w_j \cdot |x'_j - x_j| \right]$$

gdzie:

- $L(f(x'), y_{target})$ - funkcja straty
- $d(x, x')$ - odległość między instancjami
- λ - parametr balansujący
- w_j - wagi dla poszczególnych cech

Przykład z aplikacją kredytową:

- **Obecna sytuacja:** Kredyt odrzucony
- **Pytanie:** Co muszę zmienić, aby otrzymać kredyt?
- **Counterfactual:** "Gdyby Twój roczny dochód był o 5000 PLN wyższy ORAZ gdybyś miał o 2 lata dłuższą historię kredytową, Twój kredyt zostałby zatwierdzony"

WYKŁAD 7 – COUNTERFACTUALS ALGORYTM

Algorytm generowania Counterfactuals (Wachter et al., 2017)

1. **Inicjalizacja:** Zacznij od oryginalnej instancji $x^{(0)} = x$
2. **Dla każdej iteracji** $t = 1, 2, \ldots, T$**:**
   - Oblicz gradient funkcji celu:

$$\nabla_x \mathcal{L}(x^{(t)}) = \nabla_x L(f(x^{(t)}), y_{target}) + \lambda \nabla_x d(x, x^{(t)})$$

   - Aktualizuj kandydata:

$$x^{(t+1)} = x^{(t)} - \alpha \nabla_x \mathcal{L}(x^{(t)})$$

   - Zastosuj ograniczenia (feasibility, immutability, categorical)
   - Jeśli $f(x^{(t+1)}) = y_{target}$ i jakość rozwiązania jest zadowalająca, STOP
3. **Zwróć:** $x^* = x^{(T)}$ jako counterfactual

Parametry

- α - learning rate (krok gradientu)
- λ - waga odległości (proximity)
- T - maksymalna liczba iteracji

WYKŁAD 7 – COUNTERFACTUALS RÓŻNORODNOŚĆ WYJAŚNIEŃ

Dlaczego wiele counterfactuals?

- Różni użytkownicy mogą preferować różne typy zmian
- Niektóre zmiany mogą być łatwiejsze do zrealizowania niż inne
- Zapewnienie większej elastyczności i zaufania do systemu

Diverse Counterfactual Explanations (DiCE)

Optymalizuj zbiór k counterfactuals $\{x'_1, x'_2, \ldots, x'_k\}$:

$$\min_{\{x'_i\}} \sum_{i=1}^{k} \left[L(f(x'_i), y_{target}) + \lambda_1 \cdot d(x, x'_i) \right] - \lambda_2 \cdot \text{diversity}(\{x'_i\})$$

Gdzie diversity można zmierzyć jako:

$$\text{diversity}(\{x'_i\}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} d(x'_i, x'_j)$$

WYKŁAD 7 – COUNTERFACTUALS RÓŻNORODNOŚĆ WYJAŚNIEŃ

Przykład - różne counterfactuals dla aplikacji kredytowej:

1. **CF1: Zwiększ dochód o 5000 PLN**
2. **CF2: Zmniejsz zadłużenie o 3000 PLN i zwiększ staż pracy o 1 rok**
3. **CF3: Zwiększ oszczędności o 10000 PLN i popraw credit score o 50 punktów**

WYKŁAD 7 – COUNTERFACTUALS

Zalety:

- ✓ Bardzo intuicyjne dla użytkowników końcowych
- ✓ Dostarczają actionable insights (konkretne działania)
- ✓ Model-agnostic (działa z każdym modelem)
- ✓ Naturalne dla człowieka ("co gdyby...?")
- ✓ Mogą uwzględniać ograniczenia domenowe
- ✓ Zgodne z GDPR (prawo do wyjaśnienia)

Wady:

- ✗ Kosztowne obliczeniowo (optymalizacja dla każdej instancji)
- ✗ Mogą sugerować nierealistyczne zmiany
- ✗ Brak gwarancji znalezienia rozwiązania
- ✗ Wrażliwe na dobór parametrów (λ, wagi cech)
- ✗ Nie wyjaśniają globalnego zachowania modelu
- ✗ Mogą istnieć nieskończenie wiele counterfactuals

WYKŁAD 7 – COUNTERFACTUALS ZASTOSOWANIA PRAKTYCZNE

Finanse: Wyjaśnienie decyzji kredytowych

Medycyna: Alternatywne ścieżki leczenia

HR: Warunki akceptacji kandydata do pracy

Marketing: Optymalizacja profilu klienta

WYKŁAD 7 – ANCHORS WPROWADZENIE

Definicja

Anchor (reguła kotwicząca) to zbiór warunków (reguł decyzyjnych) dla instancji, które są wystarczające do "zakotwiczenia" predykcji, niezależnie od wartości innych cech.

Intuicja

"Jakie warunki muszą być spełnione, aby predykcja była stabilna?"

Przykład z klasyfikacją sentymentu recenzji filmu

- Predykcja: Pozytywna recenzja (95% pewności)
- Anchor:
  - IF "excellent" IN recenzja AND
  - IF "brilliant" IN recenzja
  - THEN predykcja = Pozytywna (z precision = 0.95)
- Interpretacja: Jeśli recenzja zawiera słowa "excellent" i "brilliant", model z 95% pewnością klasyfikuje ją jako pozytywną, niezależnie od innych słów w tekście

WYKŁAD 7 – ANCHORS

Definicja

Anchor A dla instancji x i predykcji $f(x)$ to reguła, która spełnia:

$$\mathbb{E}_{\mathcal{D}(z|A)}\left[\mathbb{1}_{f(x)=f(z)}\right] \geq \tau$$

Gdzie:

- $\mathcal{D}(z|A)$ - rozkład instancji z spełniających warunek anchor A
- $\mathbb{1}_{f(x)=f(z)}$ - funkcja wskaźnikowa (1 jeśli predykcje są takie same, 0 w przeciwnym wypadku)
- τ - próg precyzji (np. 0.95)

Coverage (pokrycie)

$$\mathrm{cov}(A) = \mathbb{E}_{\mathcal{D}(z)}\left[\mathbb{1}_{A(z)}\right]$$

Odsetek instancji w zbiorze danych, dla których anchor A jest spełniony

Problem optymalizacyjny

WYKŁAD 7 – ANCHORS
PROBLEM OPTYMALIZAYCJNY

Problem optymalizacyjny

Znajdź anchor A^* maksymalizujący coverage przy ograniczeniu precision:

$$A^* = \arg\max_{A} \mathrm{cov}(A) \quad \text{s.t.} \quad \mathrm{prec}(A) \geq \tau$$

WYKŁAD 7 – ANCHORS ALGORYTM

Algorytm Anchors (Ribeiro et al., 2018)

Główna idea: Używamy beam search do eksploracji przestrzeni reguł

1. **Inicjalizacja:**
   - Zacznij z pustym anchor $A = \emptyset$
   - Kandydaci = wszystkie możliwe predykaty (warunki) dla instancji x
2. **Beam Search:**
   - Dla każdego kandydata A' w beam:
     - Generuj perturbacje spełniające A' poprzez próbkowanie $z \sim \mathcal{D}(z|A')$
     - Oblicz precision: $\mathbf{prec}(A') = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(x)=f(z_i)}$
     - Oblicz coverage: $\mathbf{cov}(A')$
     - Jeśli $\mathbf{prec}(A') \geq \tau$, zapisz A' jako kandydata
   - Rozszerz najlepsze anchors o kolejne predykaty
3. **Zwróć:** Anchor z najwyższym coverage spełniający próg precision

WYKŁAD 7 – ANCHORS ALGORYTM

Multi-Armed Bandit dla efektywności

Algorytm używa MAB (KL-LUCB) do efektywnego oszacowania precision bez próbkowania wszystkich możliwych perturbacji.

WYKŁAD 7 – ANCHORS PRZYKŁAD

Problem: Klasyfikacja ryzyka kredytowego

Instancja:

- Wiek: 35 lat
- Dochód: 75,000 PLN
- Stosunek zadłużenia do dochodu: 25%
- Credit score: 720
- Staż pracy: 5 lat

Predykcja modelu: NISKIE RYZYKO

Znaleziony Anchor (precision = 0.96, coverage = 0.35):

IF Credit score ≥ 700 **AND** Stosunek zadłużenia do dochodu ≤ 30%

THEN Predykcja = NISKIE RYZYKO

Interpretacja

- Dla 96% klientów spełniających te dwa warunki, model przewiduje niskie ryzyko
- Te dwa warunki są wystarczające - inne cechy (wiek, dochód, staż) nie mają znaczenia
- 35% klientów w zbiorze danych spełnia te warunki

WYKŁAD 7 – ANCHORS

Zalety:

- ✓ Proste, interpretatywne reguły if-then
- ✓ Model-agnostic
- ✓ Uwzględniają interakcje między cechami
- ✓ Precyzja i pokrycie są łatwe do interpretacji
- ✓ Stabilne wyjaśnienia (robustness)
- ✓ Dobrze działają dla danych tabelarycznych, tekstowych i obrazów

Wady:

- ✗ Kosztowne obliczeniowo (wymaga wielu predykcji)
- ✗ Mogą nie istnieć anchors o wysokim coverage
- ✗ Wybór progu precision τ jest arbitralny
- ✗ Nie dostarczają informacji o kierunku wpływu
- ✗ Mogą tworzyć bardzo długie reguły (niska prostota)
- ✗ Wyjaśniają tylko lokalne zachowanie

WYKŁAD 7 – ANCHORS VS LIME

Właściwość	LIME	Anchors
Typ wyjaśnienia	Model liniowy (wagi cech)	Reguły if-then
Interpretacja	Trudniejsza (współczynniki)	Łatwiejsza (reguły)
Stabilność	Niska (wrażliwe na perturbacje)	Wysoka (threshold-based)
Coverage	Tylko lokalne sąsiedztwo	Mierzalne i kontrolowane

WYKŁAD 7 – PROTOTYPES WPROWADZENIE

Definicja

Prototypes to reprezentatywne przykłady z danych treningowych, które najlepiej charakteryzują określone klasy lub regiony przestrzeni cech. Są to "typowe" instancje, które model używa do podejmowania decyzji.

WYKŁAD 7 – PROTOTYPES WPROWADZENIE

Typy metod opartych na przykładach

1. Prototypes

Pytanie: "Jakie przykłady najlepiej reprezentują daną klasę?"

- Typowe instancje klasy
- Reprezentatywne cechy
- Centroidy klastrów

2. Criticisms (Krytyki)

Pytanie: "Jakie przykłady są nietypowe lub źle reprezentowane?"

- Outliers
- Boundary cases
- Trudne przykłady

3. Influential Instances (Wpływowe przykłady)

Pytanie: "Które przykłady treningowe miały największy wpływ na tę predykcję?"

- Influence functions
- TracIn
- Representer Point Selection

WYKŁAD 7 – PROTOTYPES MMD-CRITIC

Maximum Mean Discrepancy (MMD)

MMD mierzy różnicę między dwoma rozkładami prawdopodobieństwa P i Q w przestrzeni RKHS:

$$\mathrm{MMD}^2(P, Q) = \mathbb{E}_{x,x' \sim P}[k(x, x')] + \mathbb{E}_{y,y' \sim Q}[k(y, y')] - 2\mathbb{E}_{x \sim P, y \sim Q}[k(x, y)]$$

gdzie $k(\cdot, \cdot)$ jest funkcją jądra (np. RBF)

WYKŁAD 7 – PROTOTYPES
ALGORYTM MMD-CRITIC (KIM ET AL., 2016)

Faza 1 Wybór Prototypes

Znajdź podzbiór $S \subset \mathcal{D}$ maksymalizujący:

$$\arg \max_{S, |S|=m} \text{MMD}^2(\mathcal{D}, S)$$

Greedy selection: Iteracyjnie dodawaj instancję maksymalizującą wzrost MMD

Faza 2 Wybór Criticisms

Znajdź podzbiór $C \subset \mathcal{D} \setminus S$ maksymalizujący:

$$\arg \max_{C, |C|=n} \text{witness}(C|S)$$

gdzie witness function identyfikuje regiony słabo pokryte przez prototypes

WYKŁAD 7 – PROTOTYPES INFLUENCE FUNCTIONS

Pytanie:

"Który przykład treningowy miał największy wpływ na predykcję dla konkretnej instancji testowej?"

Influence Function (Koh & Liang, 2017)

Wpływ usunięcia instancji treningowej z na stratę dla instancji testowej z_{test}:

$$\mathcal{I}(z, z_{test}) = -\nabla_\theta L(z_{test}, \hat{\theta})^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta})$$

Gdzie:

- $\hat{\theta}$ - parametry wytrenowanego modelu
- $L(\cdot, \theta)$ - funkcja straty
- $H_{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta})$ - macierz Hessego

Interpretacja

- Dodatnia wartość $\mathcal{I}(z, z_{test}) > 0$: Usunięcie z zwiększyłoby stratę dla z_{test} (pozytywny wpływ)
- Ujemna wartość $\mathcal{I}(z, z_{test}) < 0$: Usunięcie z zmniejszyłoby stratę dla z_{test} (negatywny wpływ)

WYKŁAD 7 – PROTOTYPES ZASTOSOWANIE W PRAKTYCE

- **Debugging:** Identyfikacja problemów z danymi treningowymi
- **Data auditing:** Wykrywanie błędnych etykiet
- **Zrozumienie modelu:** Jakie przykłady model uważa za typowe
- **Aktywne uczenie:** Wybór przykładów do etykietowania

WYKŁAD 7 – PROTOTYPES

Zalety:

- ✓ Bardzo intuicyjne (przykłady rzeczywiste)
- ✓ Nie wymagają znajomości cech/modelu
- ✓ Pomagają zrozumieć dane treningowe
- ✓ Użyteczne w debugowaniu modeli
- ✓ Naturalne dla ludzkiej percepcji (case-based reasoning)
- ✓ Mogą wykryć problemy z danymi

Wady:

- ✗ Nie wyjaśniają dlaczego instancje są podobne
- ✗ Wrażliwe na wybór metryki odległości
- ✗ Kosztowne dla dużych zbiorów danych
- ✗ Influence functions: wymagają odwrotności Hessjanu ($O(p^3)$)
- ✗ Prototypes mogą nie być reprezentatywne przy złożonych danych
- ✗ Nie dostarczają informacji o ważności cech

WYKŁAD 7 – PROTOTYPES

Warianty metod Prototype

Metoda	Opis	Złożoność
k-NN	k najbliższych sąsiadów	O(n) per query
MMD-critic	Prototypes + Criticisms	$O(n^2m)$
Influence Functions	Wpływ treningowy na test	$O(np^2 + p^3)$
TracIn	Tracing influence (gradients)	O(nT) gdzie T=#epochs

WYKŁAD 7 – METODY KOHORTOWE WPROWADZENIE

Definicja

Metody kohortowe (Cohort Methods) analizują zachowanie modelu dla grup podobnych instancji, zamiast pojedynczych obserwacji lub całego modelu globalnie.

Motywacja

- **Globalne metody** (PDP, Feature Importance) mogą maskować heterogeniczność - różne grupy mogą reagować inaczej
- **Lokalne metody** (LIME, SHAP local) nie pokazują wzorców w grupach
- **Metody kohortowe wypełniają tą lukę: pokazują jak model zachowuje się dla różnych segmentów populacji**

WYKŁAD 7 – METODY KOHORTOWE WPROWADZENIE

Trzy główne podejścia

Subgroup Analysis

Analiza zachowania modelu w predefiniowanych grupach

Segment-wise PDP

Partial Dependence dla automatycznie wykrytych segmentów

Group SHAP

Agregacja wartości SHAP w grupach

WYKŁAD 7 – METODY KOHORTOWE SUBGROUP ANALYSIS

Definicja

Analiza wydajności i zachowania modelu w predefiniowanych lub odkrytych podgrupach danych.

Typy subgroup analysis:

1. **A priori** (z góry zdefiniowane grupy):
   - Demografia: Wiek (młodzi/starsi), płeć, region
   - Kliniczna: Ciężkość choroby, comorbidities
   - Biznesowa: Nowi/stali klienci, segment rynku
2. **Data-driven** (odkrywane z danych):
   - Klasteryzacja instancji
   - Decision tree subgroups
   - ICE curve clustering

WYKŁAD 7 – METODY KOHORTOWE SUBGROUP ANALYSIS

Metryki do porównania

Dla każdej grupy G_k:

$$\begin{aligned} \text{Performance}_k &= \text{Accuracy}(G_k), \text{AUC}(G_k), \dots \\ \text{Feature Importance}_k &= \text{PI}(\text{feature}_j, G_k) \\ \text{Fairness metrics}_k &= \text{Disparate Impact}, \text{Equal Opportunity} \end{aligned}$$

WYKŁAD 7 – METODY KOHORTOWE
SUBGROUP ANALYSIS PRZYKŁAD

Problem: Model predykcji ryzyka sercowo-naczyniowego

Grupy analizowane:

G1: Wiek < 50 lat

G2: Wiek 50-65 lat

G3: Wiek > 65 lat

Wyniki analizy:

Grupa	AUC	Top Feature	Feature Importance
G1 (< 50)	0.82	Cholesterol	0.35
G2 (50-65)	0.88	Ciśnienie krwi	0.42
G3 (> 65)	0.76	Historia rodzinna	0.38

WYKŁAD 7 – METODY KOHORTOWE
SUBGROUP ANALYSIS PRZYKŁAD

Wnioski kliniczne

- **Młodsi pacjenci:** Model opiera się głównie na poziomie cholesterolu
- **Wiek średni:** Ciśnienie krwi jest głównym predyktorem
- **Starsi pacjenci:** Model ma niższą dokładność, historia rodzinna staje się ważniejsza
- **Rekomendacja: Model może wymagać rekalibracji dla grupy G3**

Grupa	AUC	Top Feature	Feature Importance
G1 (< 50)	0.82	Cholesterol	0.35
G2 (50-65)	0.88	Ciśnienie krwi	0.42
G3 (> 65)	0.76	Historia rodzinna	0.38

WYKŁAD 7 – METODY KOHORTOWE SEGMENT WISE PDP

Motywacja

Globalny PDP pokazuje średni efekt cechy, ale może maskować heterogeniczne efekty w różnych segmentach populacji.

WYKŁAD 7 – METODY KOHORTOWE
SEGMENT WISE PDP ALGORYTM

1. **Oblicz ICE curves** dla wszystkich instancji $i = 1, \ldots, n$:

$$\mathrm{ICE}_i(x_j) = f(x_j, x_{\setminus j}^{(i)})$$

2. **Klasteryzacja ICE curves:**
   - Użyj k-means, hierarchical clustering lub innej metody
   - Metryka odległości: np. Dynamic Time Warping, euclidean
3. **Dla każdego klastra C_k:**

$$\mathrm{PDP}_{C_k}(x_j) = \frac{1}{|C_k|} \sum_{i \in C_k} \mathrm{ICE}_i(x_j)$$

4. **Wizualizuj** segment-wise PDP curves dla każdego klastra

WYKŁAD 7 – METODY KOHORTOWE
SEGMENT WISE PDP

Interpretacja:

- **Każdy klaster reprezentuje segment populacji o podobnej reakcji na zmianę cechy**
- **Różnice między PDP klastrów pokazują heterogeniczność efektu**
- **Można zidentyfikować subpopulacje z różnymi wzorcami zachowań**

WYKŁAD 7 – METODY KOHORTOWE
GROUP SHAP

Motywacja

SHAP values dostarczają lokalnych wyjaśnień dla pojedynczych instancji. Group SHAP agreguje te wartości dla grup, pozwalając zrozumieć typowe wzorce wyjaśnień w kohortach.

WYKŁAD 7 – METODY KOHORTOWE
GROUP SHAP

Definicja

Dla grupy instancji $G = \{x^{(1)}, \ldots, x^{(m)}\}$, średnia wartość SHAP dla cechy j:

$$\text{GroupSHAP}_G(j) = \frac{1}{m} \sum_{i=1}^{m} \phi_j(x^{(i)})$$

Można również obliczyć:

- **Mediana:** $\text{median}(\{\phi_j(x^{(i)})\}_{i=1}^{m})$
- **Odchylenie standardowe:** $\text{std}(\{\phi_j(x^{(i)})\}_{i=1}^{m})$
- **Rozkład:** Histogram wartości SHAP w grupie

WYKŁAD 7 – METODY KOHORTOWE GROUP SHAP

Warianty agregacji

1. Mean Absolute SHAP

$$\text{GroupSHAP}_G^{\text{abs}}(j) = \frac{1}{m}\sum_{i=1}^{m}|\phi_j(x^{(i)})|$$

Pokazuje średnią siłę wpływu (niezależnie od kierunku)

2. Directional Agreement

$$\text{Agreement}_G(j) = \frac{|\sum_{i=1}^{m}\text{sign}(\phi_j(x^{(i)}))|}{m}$$

Mierzy konsensus co do kierunku wpływu (0-1)

WYKŁAD 7 – METODY KOHORTOWE
GROUP SHAP PRZYKŁAD

Problem: Model predykcji churn (odejście klienta)

Grupy analizowane:

- G1: Nowi klienci (< 6 miesięcy)
- G2: Klienci lojalni (> 2 lata)

Group SHAP Analysis

Cecha	G1: Nowi (mean SHAP)	G2: Lojalni (mean SHAP)	Różnica
Cena usługi	**+0.35**	+0.08	★★★
Jakość obsługi	-0.15	**-0.42**	★★
Liczba kontaktów	-0.05	-0.12	★
Program lojalnościowy	-0.02	**-0.38**	★★★

WYKŁAD 7 – METODY KOHORTOWE
GROUP SHAP PRZYKŁAD

Interpretacja i wnioski:

- **Nowi klienci:** Bardzo wrażliwi na cenę (+0.35 SHAP), mniej na jakość obsługi
- **Lojalni klienci:** Jakość obsługi (-0.42) i program lojalnościowy (-0.38) są kluczowe
- **Rekomendacja:**
  - **Nowi**: Oferty cenowe, elastyczne plany
  - **Lojalni**: Inwestować w jakość obsługi i benefity lojalnościowe

Cecha	G1: Nowi (mean SHAP)	G2: Lojalni (mean SHAP)	Różnica
Cena usługi	**+0.35**	+0.08	★★★
Jakość obsługi	-0.15	**-0.42**	★★
Liczba kontaktów	-0.05	-0.12	★
Program lojalnościowy	-0.02	**-0.38**	★★★

WYKŁAD 7 – METODY KOHORTOWE PORÓWNANIE

Metoda	Typ analizy	Grupy	Output	Zastosowanie
Subgroup Analysis	Model performance	Predefiniowane	Metryki per grupa	Fairness, debugging
Segment-wise PDP	Feature effect	Data-driven (ICE)	PDP curves per segment	Heterogeniczne efekty
Group SHAP	Feature attribution	Obie	Agregowane SHAP values	Porównanie grup

WYKŁAD 7 – METODY KOHORTOWE PORÓWNANIE

Kiedy używać której metody?

Subgroup Analysis

- ✓ Znasz istotne grupy (domena)
- ✓ Fairness / bias detection
- ✓ Regulatory requirements

Segment-wise PDP

- ✓ Odkrywanie ukrytych segmentów
- ✓ Heterogeniczne efekty cech
- ✓ Feature engineering insights

Group SHAP

- ✓ Porównanie wzorców wyjaśnień
- ✓ Feature importance per segment
- ✓ Personalizacja strategii

WYKŁAD 7 – METODY KOHORTOWE PORÓWNANIE

Best Practices

- Zawsze sprawdź **rozmiar grupy** - małe grupy mogą dawać niestabilne wyniki
- Użyj **multiple testing correction** przy porównywaniu wielu grup
- **Wizualizuj rozkłady**, nie tylko średnie
- **Łącz metody dla pełniejszego obrazu**

WYKŁAD 7 – METODY KOHORTOWE

Zalety:

- ✓ Ujawniają heterogeniczność efektów
- ✓ Pomagają w wykrywaniu bias/fairness issues
- ✓ Umożliwiają personalizację strategii
- ✓ Łączą globalne i lokalne perspektywy
- ✓ Mogą odkryć nieoczekiwane subpopulacje
- ✓ Przydatne w regulatory/auditing contexts
- ✓ Dostarczają actionable insights per segment

Wady:

- ✗ Wymagają wystarczającej liczby obserwacji per grupa
- ✗ Multiple testing problem (wiele porównań)
- ✗ Wybór grup może być arbitralny
- ✗ Data-driven segmentation może być niestabilna
- ✗ Zwiększona złożoność interpretacji
- ✗ Kosztowne obliczeniowo (k × koszt metody bazowej)

WYKŁAD 7 – PORÓWNANIE METOD

Metoda	Scope	Technique	Output	Główne zastosowanie
Counterfactuals	Local	Example-based	Alternatywne przykłady	Actionable recourse
Anchors	Local	Rule-based	Reguły if-then	Stabilne wyjaśnienia
Prototypes	Global	Example-based	Reprezentatywne przykłady	Zrozumienie klas
Influence Functions	Local	Example-based	Wpływowe przykłady	Debugging, data auditing
Subgroup Analysis	Cohort	Performance	Metryki per grupa	Fairness, bias detection
Segment-wise PDP	Cohort	Partial Dependence	PDP per segment	Heterogeniczne efekty
Group SHAP	Cohort	Feature Attribution	Agregowane SHAP	Porównanie grup

WYKŁAD 7 – PORÓWNANIE METOD

Wybór metody

1. **Pytanie:** Czy wyjaśnienie dla pojedynczej predykcji?
   - TAK → Counterfactuals (actionable) lub Anchors (stabilne)
   - NIE → Przejdź dalej
2. **Pytanie:** Czy interesują Cię konkretne grupy?
   - TAK → Metody kohortowe
   - NIE → Prototypes (reprezentacja) lub Influence (debugging)

WYKŁAD 7 – PODSUMOWANIE

1. **Counterfactuals - Best Practices**

- Zawsze definiuj **immutable features (wiek, płeć, historia)**
- Używaj **diverse counterfactuals** (DiCE) dla większej użyteczności
- Waliduj **feasibility** z ekspertami domenowymi
- Rozważ **koszty** zmian różnych cech (nie wszystkie zmiany są równie łatwe)

2. Anchors - Best Practices

- Ustaw **threshold precision** zależnie od zastosowania (regulatory: wysokie, exploratory: niższe)
- Monitoruj **coverage** - niska coverage może wskazywać na zbyt specyficzne reguły
- Używaj dla **high-stakes decisions** gdzie stabilność jest kluczowa
- Łącz z LIME/SHAP dla pełniejszego obrazu

WYKŁAD 7 – PODSUMOWANIE

3. **Prototypes - Best Practices**

- Zawsze pokazuj zarówno **prototypes JAK I criticisms**
- Użyj influence functions do **debugowania błędów** modelu
- Periodycznie sprawdzaj czy prototypes są **still representative** (data drift)
- Wizualizuj prototypes dla interpretacji (szczególnie dla obrazów/tekstu)

4**. Metody Kohortowe - Best Practices**

- Zapewnij **min. 100 obserwacji** per grupa dla stabilnych wyników
- Używaj **Bonferroni correction** lub FDR dla multiple testing
- Wizualizuj **confidence intervals**, nie tylko point estimates
- Dokumentuj **clinical/business significance**, nie tylko statistical

WYKŁAD 7 – BIBLIOTEKI PYTHON

Counterfactuals:

- **DiCE** - Diverse Counterfactual Explanations `pip install dice-ml`
- **Alibi** - Counterfactuals with RL `pip install alibi`

Anchors:

- **Alibi** - Anchor implementation `pip install alibi`
- Anchor implementacja w **anchor-exp**

Prototypes:

- **MMD-critic** - Prototypes selection
- **Captum** - Influence functions (PyTorch) `pip install captum`
- **TracIn** - Influence tracing

Metody Kohortowe:

- **scikit-learn** - Clustering dla segmentation
- **SHAP** - Group SHAP analysis
- **PDPbox** - Segment-wise PDP

ŹRÓDŁA

1. Wachter, S., Mittelstadt, B., & Russell, C. (2017). "Counterfactual Explanations Without Opening the Black Box" Harvard Journal of Law & Technology
2. Mothilal, R. K., Sharma, A., & Tan, C. (2020). "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations" FAT* 2020
3. Karimi, A. H., et al. (2021). "Model-Agnostic Counterfactual Explanations for Consequential Decisions" AISTATS 2021
4. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). "Anchors: High-Precision Model-Agnostic Explanations" AAAI 2018
5. Kim, B., Khanna, R., & Koyejo, O. (2016). "Examples are not Enough, Learn to Criticize! Criticism for Interpretability" NeurIPS 2016
6. Koh, P. W., & Liang, P. (2017). "Understanding Black-box Predictions via Influence Functions" ICML 2017
7. Pruthi, G., et al. (2020). "Estimating Training Data Influence by Tracing Gradient Descent" NeurIPS 2020
8. Goldstein, A., et al. (2015). "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation" JCGS 2015
9. Kumar, I. E., et al. (2020). "Problems with Shapley-value-based Explanations as Feature Importance Measures" ICML 2020

ŹRÓDŁA

Online

1. https://christophm.github.io/interpretable-ml-book/
2. https://shap.readthedocs.io/
3. https://shap.readthedocs.io/en/latest/overviews.html

Regulacje

1. EBA/REP/2023/28: Report on ML for IRB models
2. SR 11-7: Guidance on Model Risk Management
3. KNF: Rekomendacja P