ZASTOSOWANIE I ANALIZA METOD WYJAŚNIALNEGO UCZENIA MASZYNOWEGO NA ZBIORZE ADULT

Jan Cichowlas ◊ Ziemowit Głowaczewski ◊ Szymon Pawlonka ◊ Alicja Pruszyńska ◊ Krzysztof Stolarski

KRÓTKO O ZBIORZE

Adult

Donated on 4/30/1996

Predict whether annual income of an individual exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Social Science	Classification
Feature Type	**# Instances**	**# Features**
Categorical, Integer	48842	14

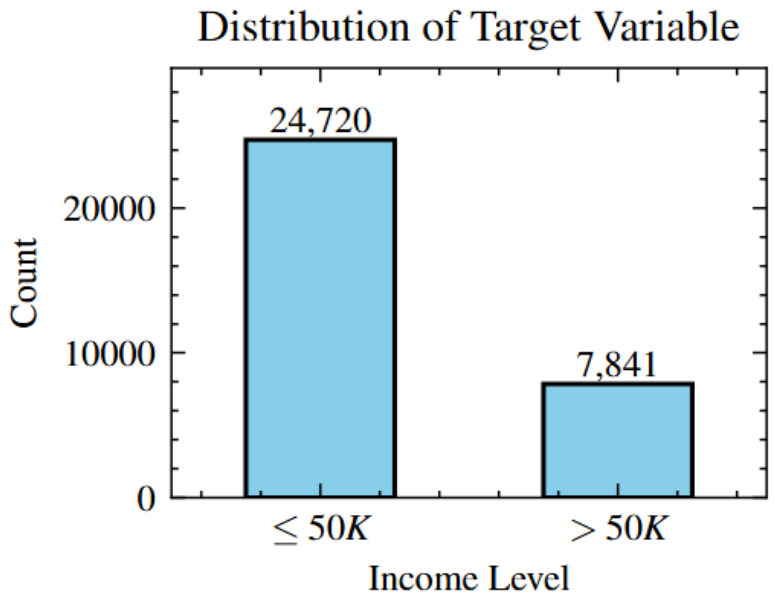
CECHY

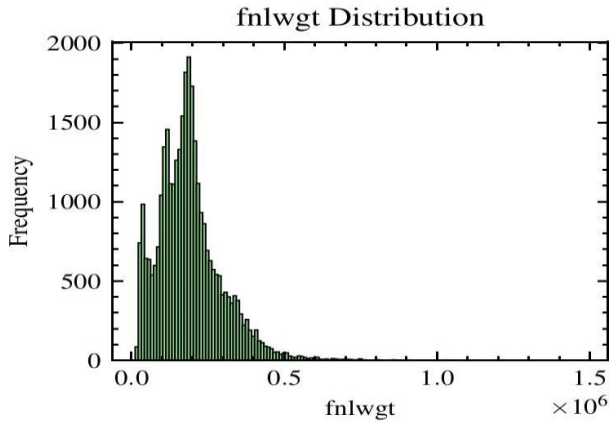
Zmienne skategoryzowane: workclass, education, marital-status, occupation, relationship, race, sex, native-country

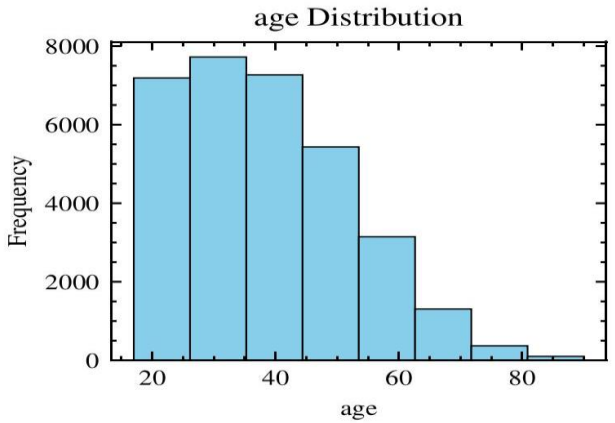
Zmienne numeryczne: age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week

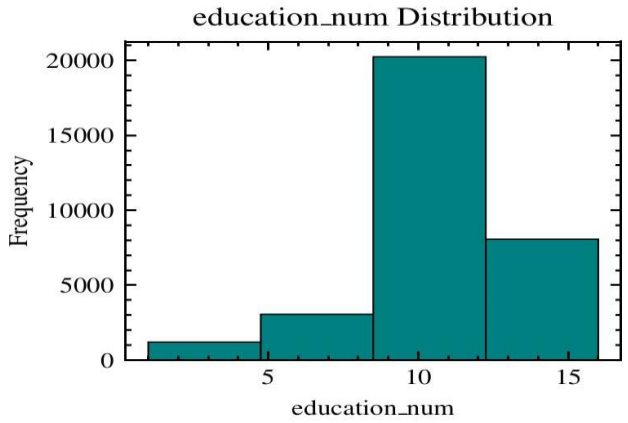
Zmienna objaśniana: income (>50K lub <=50K)

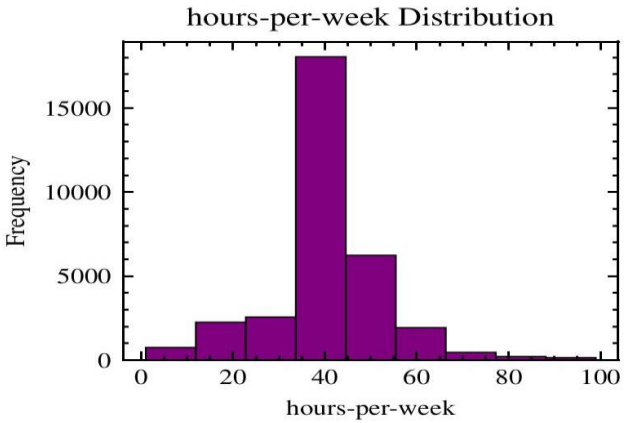
ROZKŁADY CECH

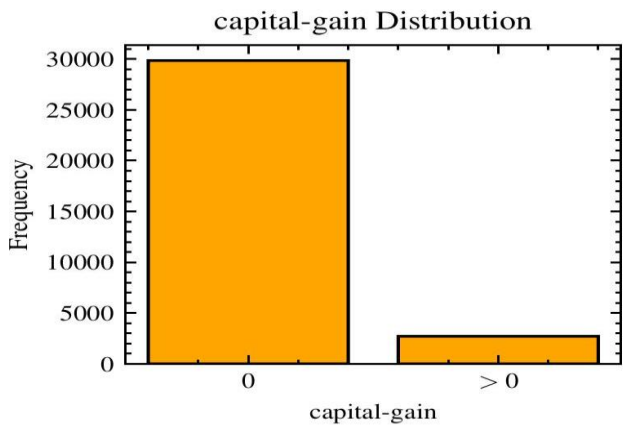


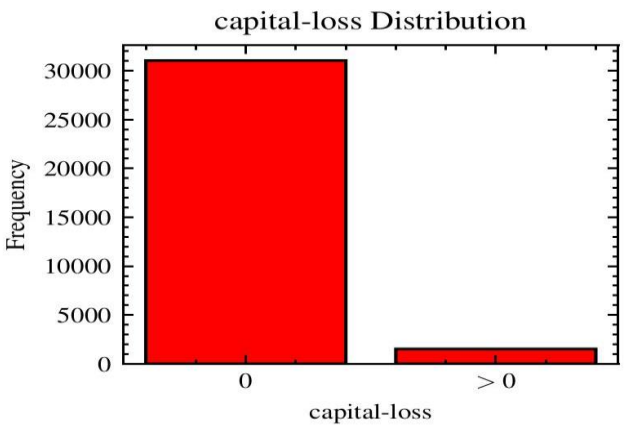












WYNIKI ANALIZOWANYCH MODELI

Tabela 1: Wyniki wybranych klasyfikatorów na zbiorze testowym przy zastosowaniu domyślnych parametrów.

Model	Accuracy	Precision	Recall
GradientBoosting	**0.87**	**0.83**	0.78
MLP	0.85	0.80	0.78
RandomForest	0.85	0.80	0.77
KNN	0.84	0.79	0.78
GaussianNB	0.82	0.76	**0.80**
SVC	0.80	0.77	0.62

METODY XAI

Oraz uzyskane przez nie wyniki

SHAP

Potrzebujemy miary, która intuicyjnie wyjaśni nam wpływ wartości zmiennej na predykcję modelu.

Dobrym pomysłem wydaje się:

$$\phi_k := \mathbb{E}[f(x) \mid x_1, \cdots, x_k] - \mathbb{E}[f(x) \mid x_1, \cdots, x_{k-1}]$$

JAK TO POLICZYĆ?

Definicja klasycznej wartości Shapley'a:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

$$= \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$$

Gdzie N - zbiór zmiennych objaśniających, i – indeks zmiennej, v – predykcja modelu dla ustalonego zbioru zmiennych

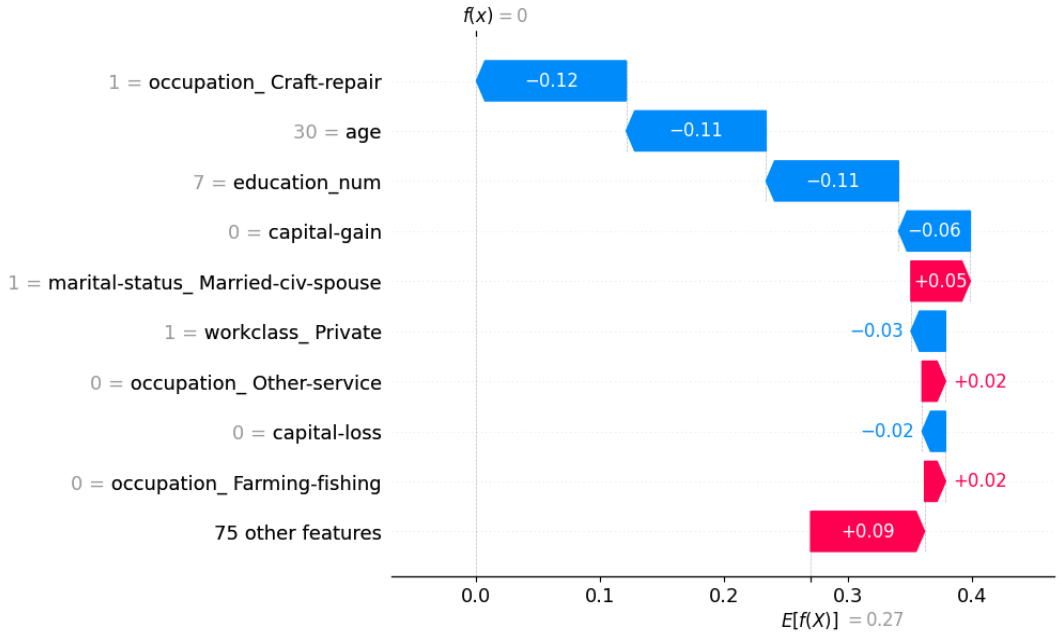
Problem: Złożoność obliczeniowa - zauważmy, że wyliczenie wartości Shapleya dla jednej zmiennej wymaga dopasowania 2**(N-1) modeli! (2**N dla wszystkich wartości)

JAK TO POLICZYĆ W SKOŃCZONYM CZASIE?

Metod jest wiele – dla większości modeli możliwe jest jedynie wyliczenie przybliżonych wartości SHAP.

Istnieją jednak przypadki, w których obliczenia da się sprowadzić do algorytmu o złożoności wielomianowej. Przykładowo - TreeSHAP.

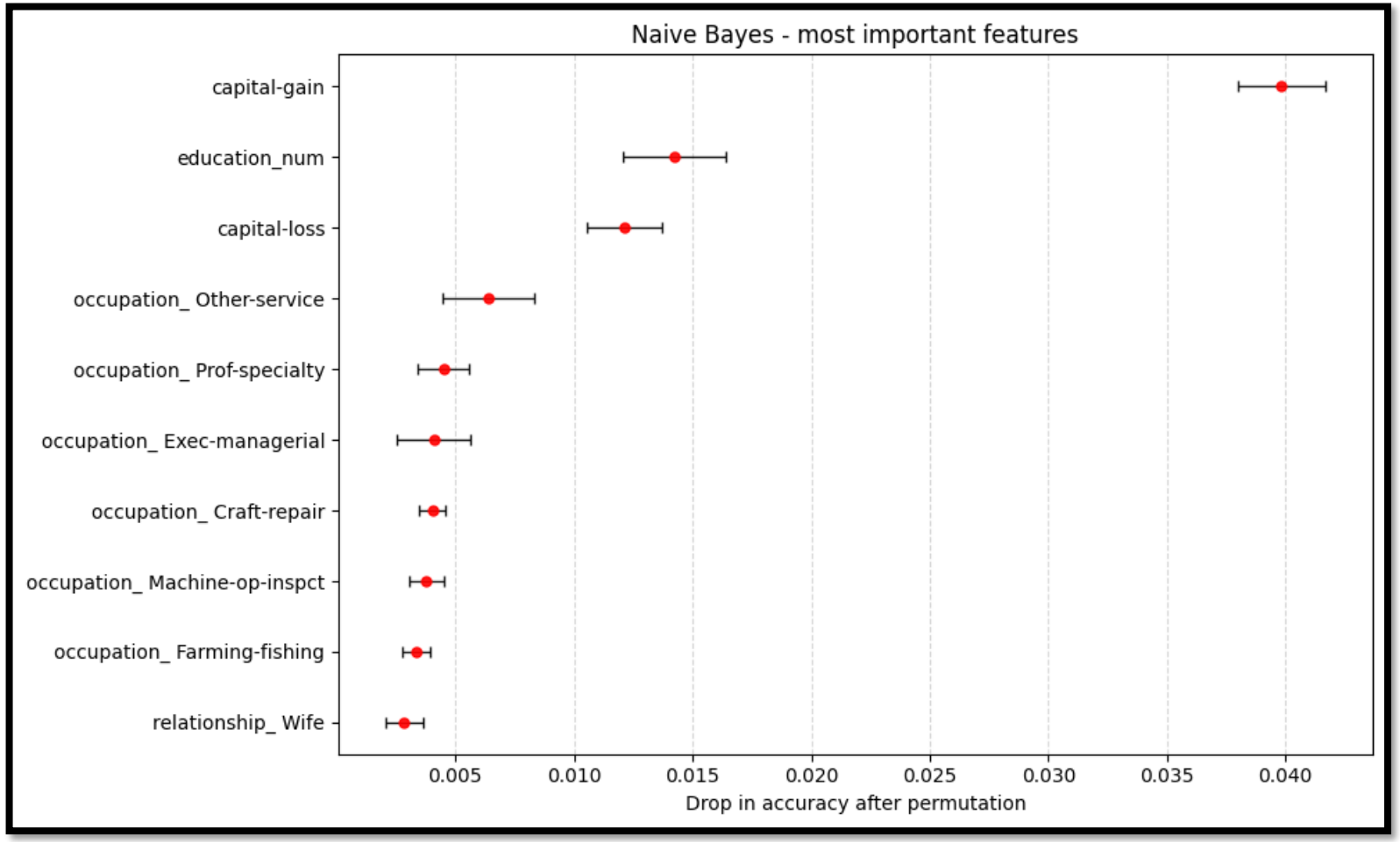
WYNIKI (DLA MLP)



PERMUTACYJNA ISTOTNOŚĆ ZMIENNYCH

Permutacyjna istotność zmiennych (ang. *permutation feature importance*) to metoda oszacowania istotności zmiennych na podstawie wzrostu błędu predykcji (lub poprawy miary dopasowania) po spermutowaniu wybranej zmiennej

$$PFI(X_j) = Acc_{\text{original}} - \mathbb{E}[Acc_{\text{permuted}}(X_j)]$$


Naive Bayes - most important features
capital-gain
education_num
capital-loss
occupation_ Other-service
occupation_ Prof-specialty
occupation_ Exec-managerial
occupation_ Craft-repair
occupation_ Machine-op-inspct
occupation_ Farming-fishing
relationship_ Wife
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
Drop in accuracy after permutation

PARTIAL DEPENDENCE PLOTS (PDP)

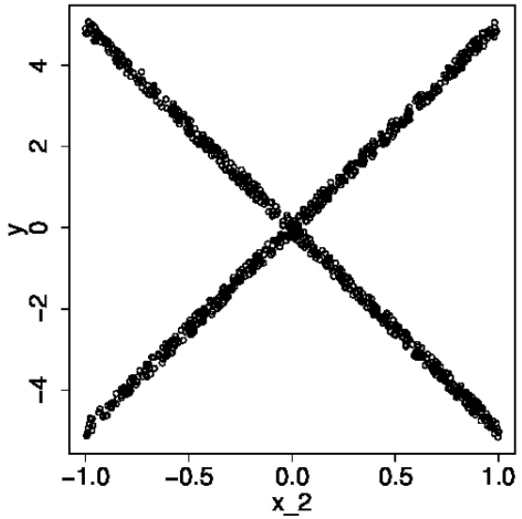
- Oznaczmy przez $S \subset \{1, \ldots p\}$ zbiór wybranych cech oraz przez C jego zbiór dopełniający
- Przez x_S rozumiemy obserwacje powstałe z $x \in X^p$ poprzez wybór cech S

$$f_S = \mathbb{E}_{\boldsymbol{x}_C}\left[f\left(\boldsymbol{x}_S, \boldsymbol{x}_C\right)\right] = \int f\left(\boldsymbol{x}_S, \boldsymbol{x}_C\right) \mathrm{dP}\left(\boldsymbol{x}_C\right)$$

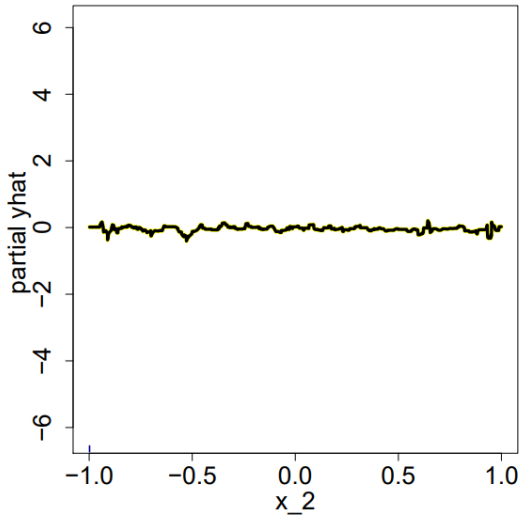
- Nie znamy f oraz nie znamy $P(\boldsymbol{x}_C)$, ale mamy nasz model i możemy obliczyć średnią

$$\hat{f}_S = \frac{1}{N} \sum_{i=1}^{N} \hat{f}\left(\boldsymbol{x}_S, \boldsymbol{x}_{Ci}\right)$$

ALE TA METODA MA ISTOTNĄ WADĘ...



(a) Scatterplot of Y versus X_2



(b) PDP

Figure 1: Scatterplot and PDP of X_2 versus Y for a sample of size 1000 from the process described in Equation 3. In this example $\hat{f}$ is fit using `SGB`. The PDP incorrectly suggests that there is no meaningful relationship between X_2 and the predicted Y.

INDIVIDUAL CONDITIONAL EXPECTATION (ICE)

- Rozważmy zbiór wszystkich próbek $\{(x_{Si}, x_{Ci})\}_{i=1}^{N}$. Krzywa ICE $\hat{f}_S^{(i)}$ przyjmuje wartości dla ustalonego x_C przy zmiennym x_S osiągającym wszystkie wartości ze zbioru treningowego

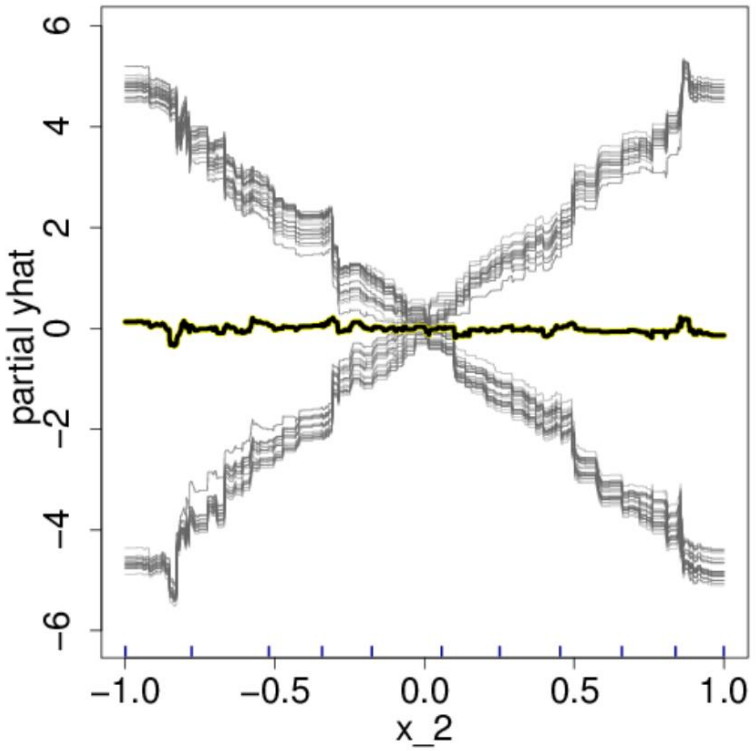
$$\hat{f}_S = \frac{1}{N} \sum_{i=1}^{N} \hat{f}\left(\boldsymbol{x}_S, \boldsymbol{x}_{Ci}\right)$$

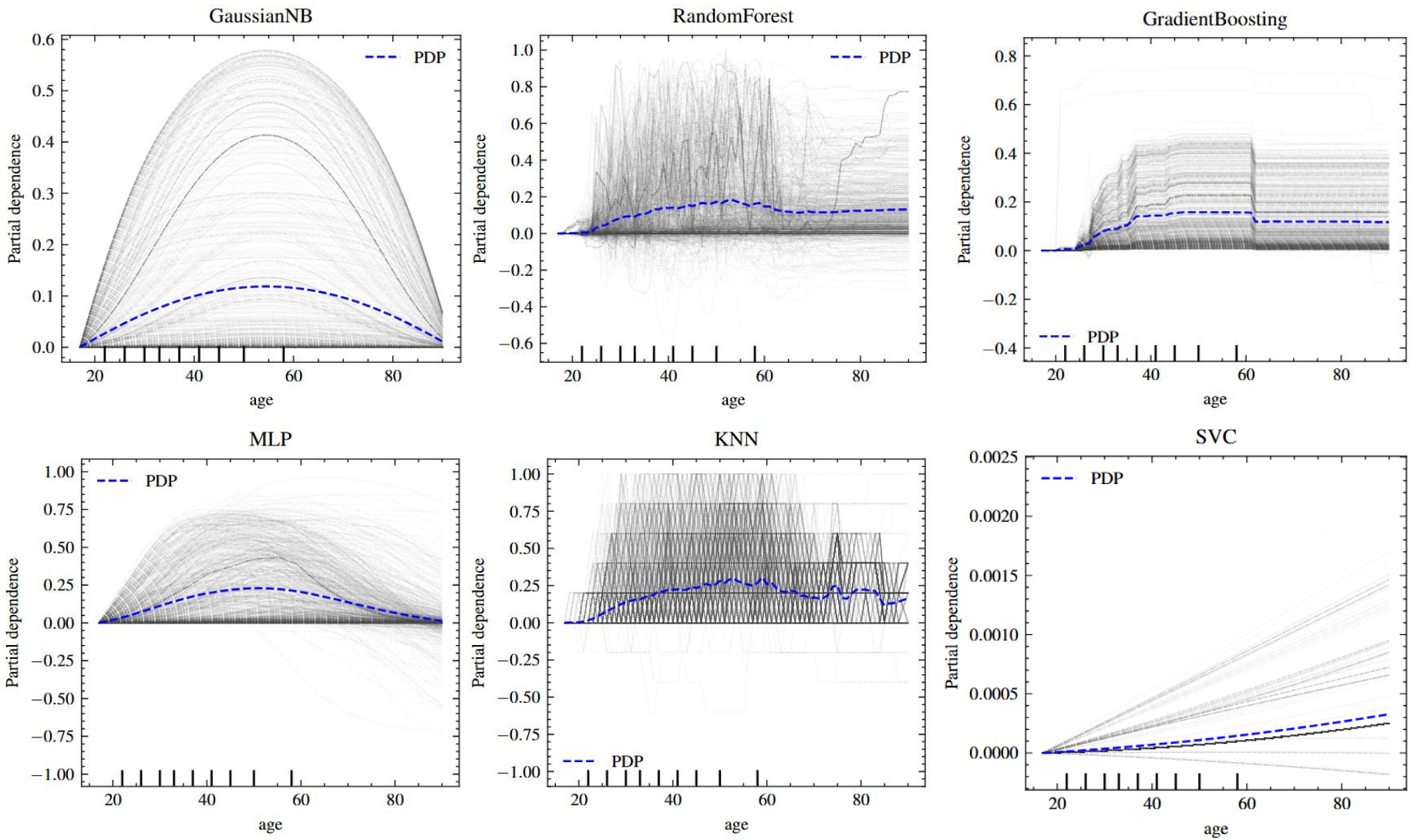
INDIVIDUAL CONDITIONAL EXPECTATION (ICE)

- Rozważmy zbiór wszystkich próbek $\{(x_{Si}, x_{Ci})\}_{i=1}^{N}$. Krzywa ICE $\hat{f}_S^{(i)}$ przyjmuje wartości dla ustalonego x_C przy zmiennym x_S osiągającym wszystkie wartości ze zbioru treningowego

$$\hat{f}_S^{(i)} = \frac{1}{N} \sum_{i=1}^{N} \hat{f}\left(\boldsymbol{x}_S, \boldsymbol{x}_{Ci}\right)$$

I TERAZ JUŻ WIDAĆ…




GaussianNB
RandomForest
GradientBoosting
MLP
KNN
SVC
PDP
Partial dependence
age

ACCUMULATED LOCAL EFFECTS (ALE)

- Załóżmy, że dysponujemy modelem predykcyjnym f i chcemy oszacować wpływ zmiennej x_j na predykcję. Niech x_{-j} oznacza wektor pozostałych zmiennych. W metodzie ALE dzielimy najpierw dziedzinę wartości, jakie x_j może przyjmować, na K przedziałów $[z_k,\ z_{k+1}]$, gdzie k = 1, 2, ..., K. Wówczas definiujemy efekt lokalny

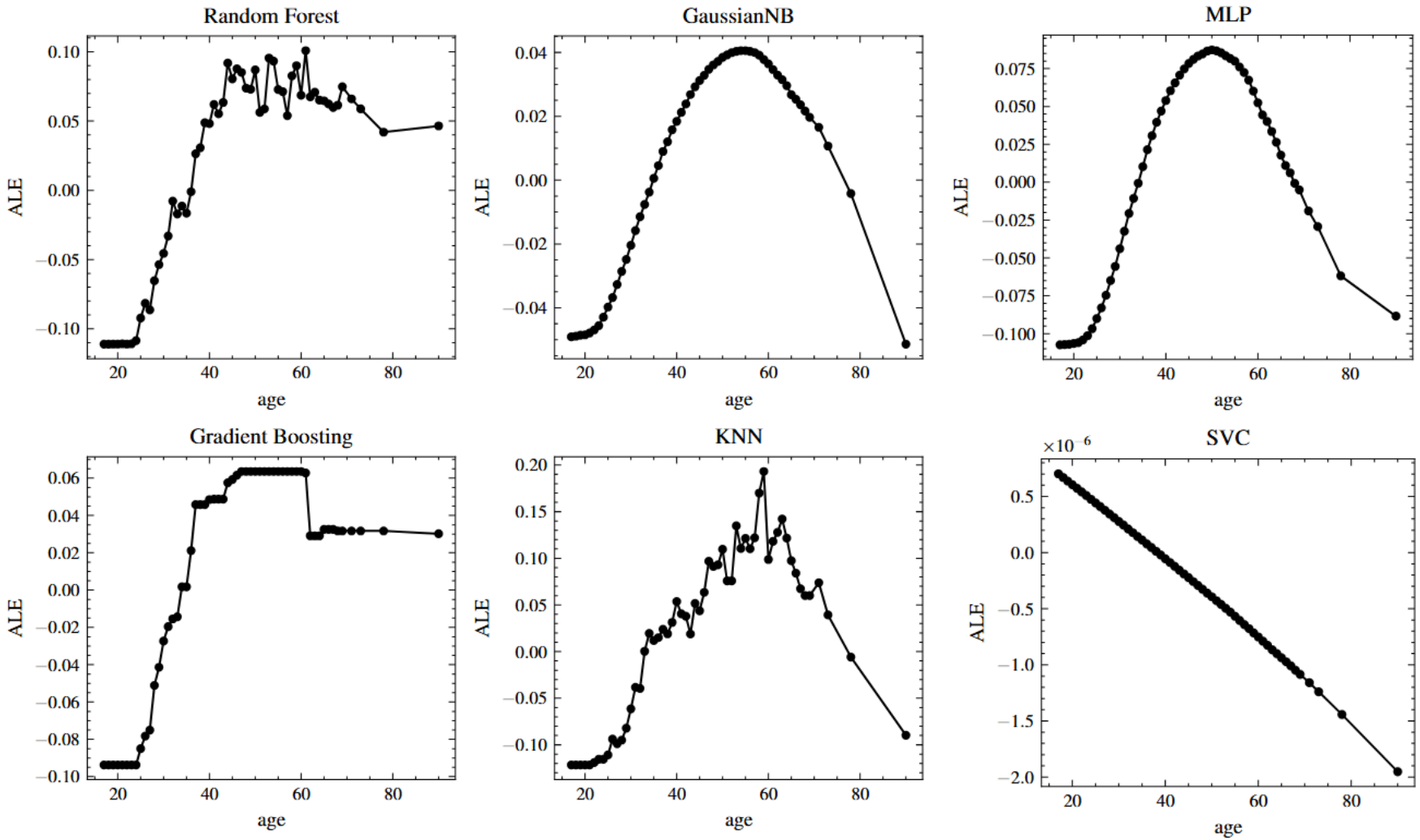
$$\Delta f_j^{(k)} = \mathbb{E}_{x_{-j}|x_j \in [z_k, z_{k+1}]} \left[f(z_{k+1}, x_{-j}) - f(z_k, x_{-j}) \right].$$

ACCUMULATED LOCAL EFFECTS (ALE)

- Efekt lokalny interpretowany jest jako średnia zmiana predykcji, gdy zmieniamy x_j z z_k do z_{k+1}, pozostawiając pozostałe cechy na wartościach występujących w danym zakresie x_j należących do $[z_k, z_{k+1}]$.

- Efekt skumulowany to suma efektów lokalnych z wszystkich przedziałów:

$$\hat{f}_{\mathrm{ALE}_j} = \sum_{l=1}^{K} \Delta f_j^{(l)}.$$


Random Forest
GaussianNB
MLP
Gradient Boosting
KNN
SVC
ALE
age
×10⁻⁶

WNIOSKI

KLUCZOWE ZMIENNE WPŁYWAJĄCE NA PREDYKCJĘ

- zysk kapitałowy – silny, pozytywny wpływ.
- poziom wykształcenia – pozytywny wpływ.
- stan cywilny – pozytywny wpływ.
- wiek– największe zarobki w średnim wieku.
- liczba godzin pracy tygodniowo – optymalnie ok. 40–50 h/tyg.

OCENA ZASTOSOWANYCH METOD XAI

- SHAP – umożliwia globalną i lokalną interpretację wpływu cech.
- ICE i PDP – dobre do wizualizacji zależności, ICE lepiej oddaje lokalne zmiany.
- ALE – preferowane przy skorelowanych cechach, np. wiek i edukacja.
- Permutacyjna istotność – łatwa do zrozumienia, potwierdza wagę najważniejszych cech.

LITERATURA

- A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking inside the black box:Visualizing statistical learning with plots of individual conditional expectation,” 2014.
- B. Becker and R. Kohavi, “Adult.” UCI Machine Learning Repository, 1996. DOI:https://doi.org/10.24432/C5XW20.

DZIĘKUJEMY

- Czy są jakieś pytania?