LIME

Emilia Majerz

15.10.2024

Local Interpretable Model-agnostic Explanations (LIME)

- Świetnie nadaje się do interpretacji modeli nauczonych na danych z dużą liczbą zmiennych (w przeciwieństwie do np. SHAPa).
- Zaproponowana w artykule “Why Should I Trust You?” Explaining the Predictions of Any Classifier (2016).
- Lokalna aproksymacja modelu "czarnej skrzynki" przez prostszy, łatwiejszy w interpretacji model.
- Nadaje się zarówno do klasyfikacji, jak i regresji.

Intuicja

- Prostszy model reprezentowany przez przerywaną linię.
- "Wyjaśnia" on zachowanie modelu czarnej skrzynki naokoło analizowanej próbki.
- Typowo wykorzystywane modele: regresja liniowa, drzewa.

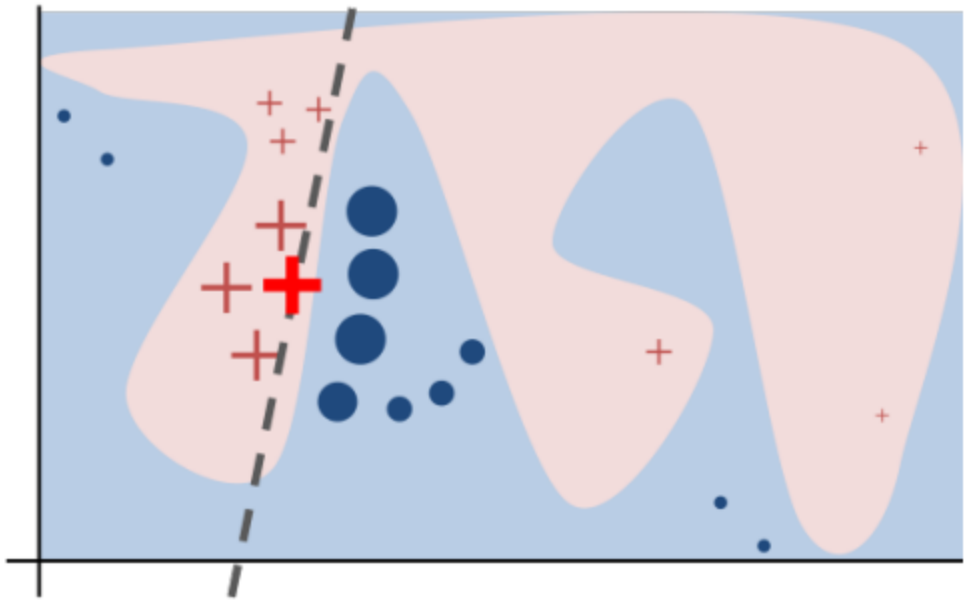


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

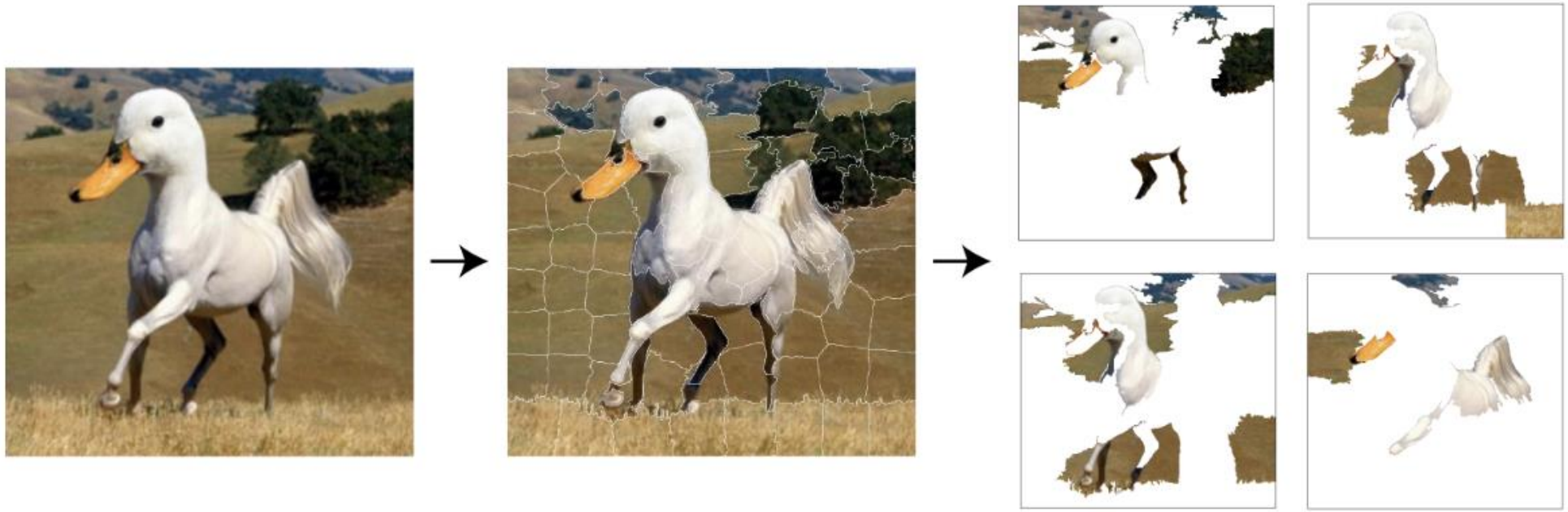
Reprezentacja danych

- Interpretowalna reprezentacja.
- Ważna szczególnie dla danych wielowymiarowych, jak obrazy.
- Piksele -> superpiksele (włączane bądź wyłączane).

Generowanie danych dla potrzeb interpretacji

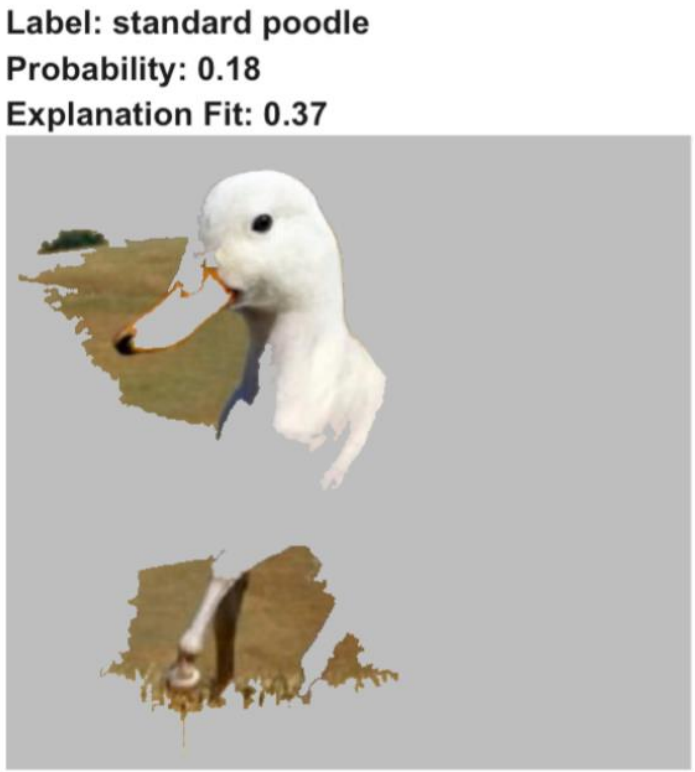
- Naokoło analizowanego przykładu.
- Nowe dane, a nie z oryginalnego zbioru.
  - W oryginalnym zbiorze może ich być za mało.
- Nowe dane: analizowany przykład poddawany perturbacjom.
  - Dane binarne: zamiana 0 na 1 i na odwrót.
  - Dane ciągłe: różne podejścia, np. dokładanie szumu, kwantyzacja i perturbacja tak uzyskanych danych binarnych.
  - Obrazy: włączanie/wyłączanie superpikseli.

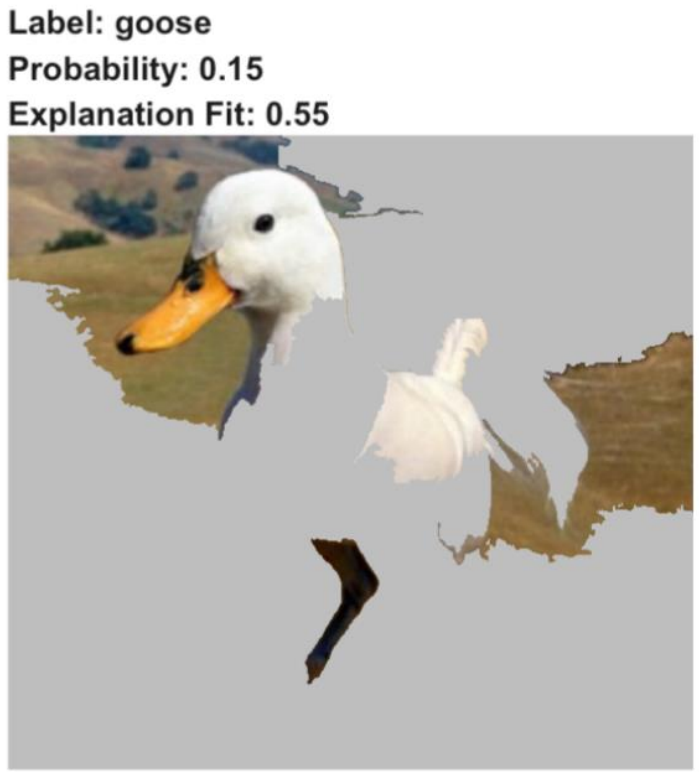
Generowanie danych – obrazy – przykład



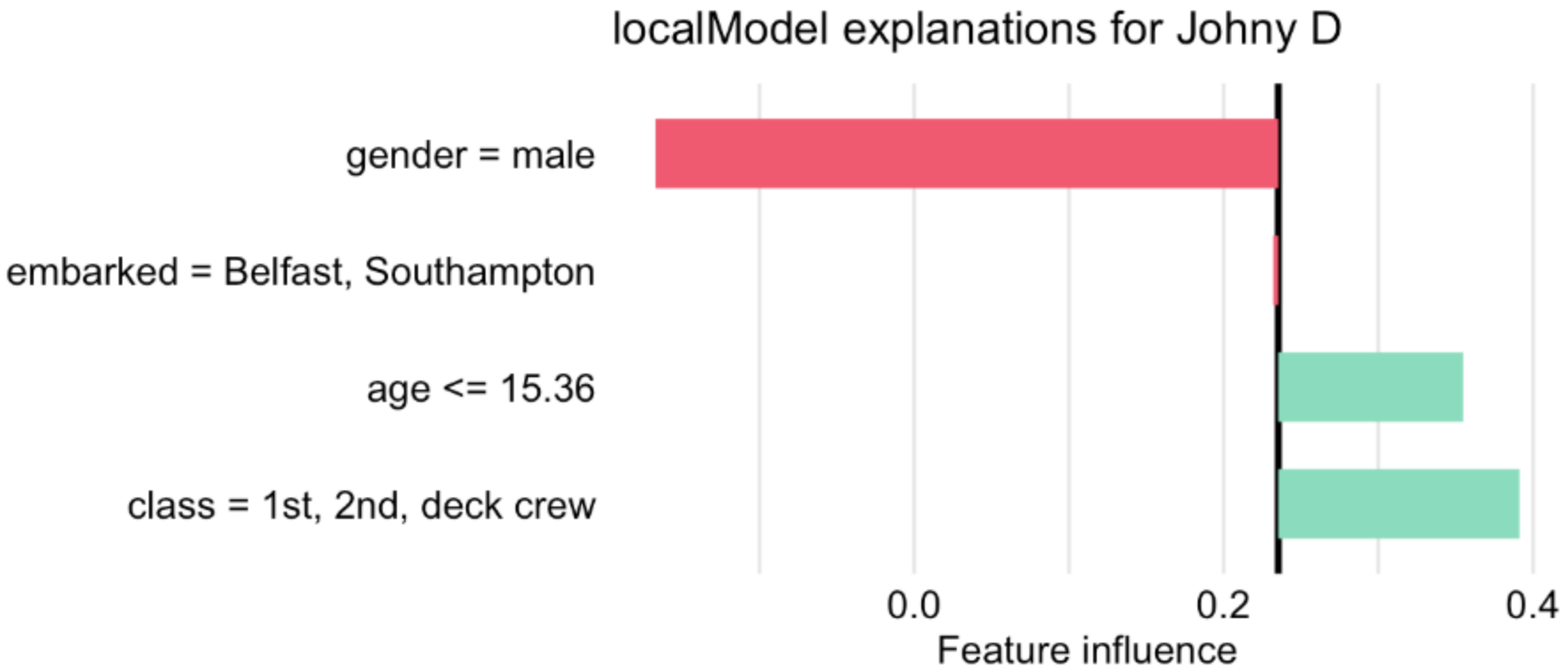
Trening lokalnego modelu

- Trening prostego (zwykle liniowego) modelu na wygenerowanych próbkach.
- Po prawej: 15 najważniejszych superpikseli dla predykcji pudla i gęsi.





Przykład - Titanic



Łatwiejsze w interpretacji dyskretne cechy.

SHAP

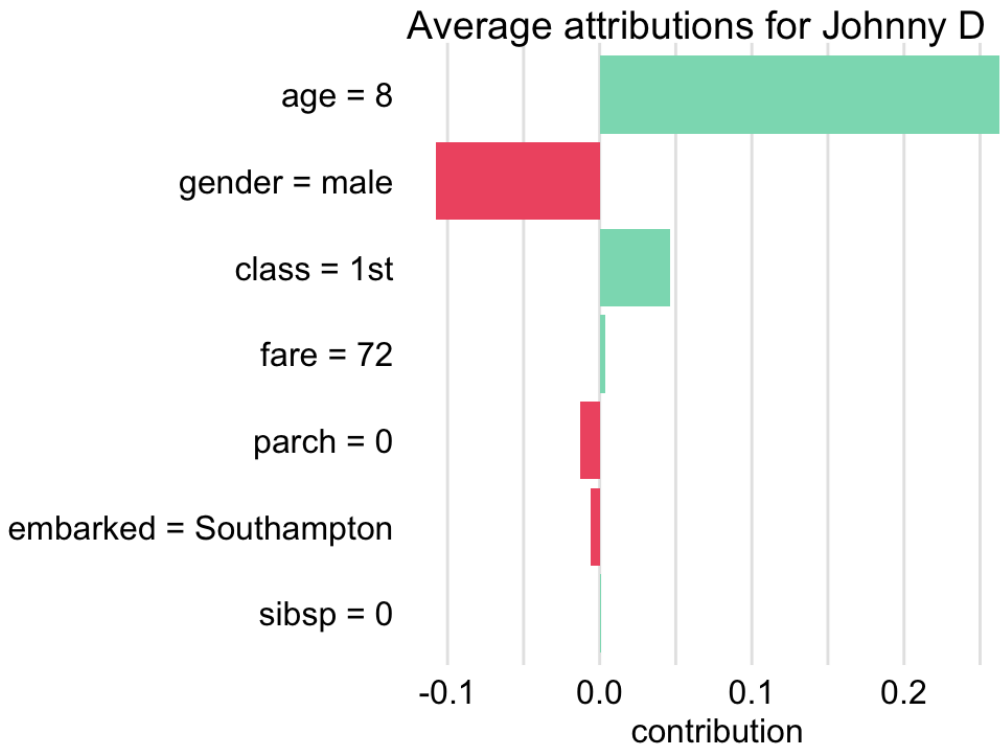
SHapley Additive exPlanations (SHAP)

- Bazuje na *wartościach Shapleya*, przedstawionych w teorii gier w 1953r.
- Wartości Shapleya rozwiązują problem sprawiedliwego podziału nagrody pomiędzy graczy, uwzględniając różnice pomiędzy graczami, a także współpracę pod-zespołów graczy.
- W świecie ML: cechy to poszczególni gracze, model - zespół, predykcja – nagroda z zawodów. SHAP pozwala odpowiedzieć na pytanie, jak podzielić predykcję modelu na wpływ poszczególnych cech.
- Przedstawiony w artykułach z 2010 i 2014 roku.
  - Štrumbelj, Erik, and Igor Kononenko. 2010. “An Efficient Explanation of Individual Classifications Using Game Theory.”
  - Štrumbelj, Erik, and Igor Kononenko. 2014. “Explaining prediction models and individual predictions with feature contributions.”

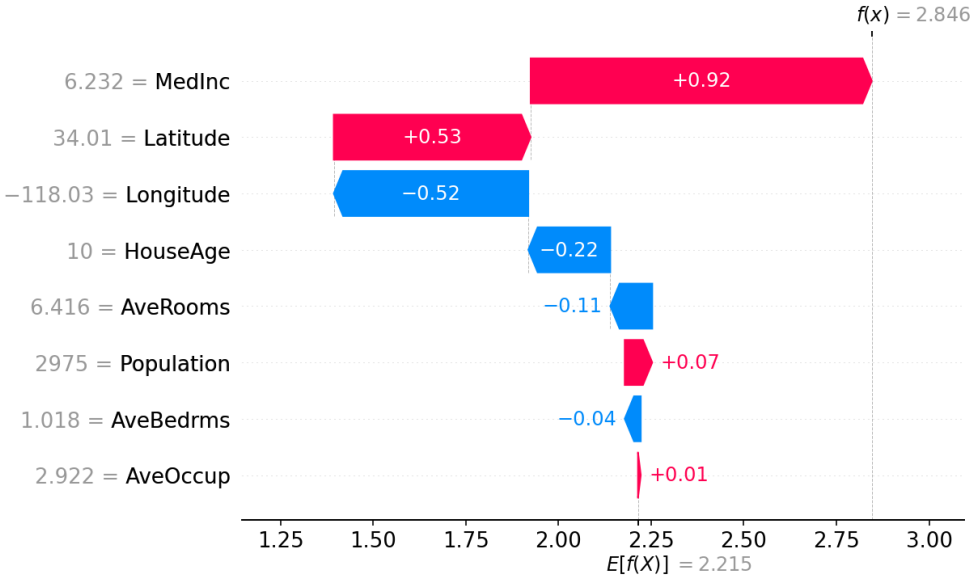
Intuicja

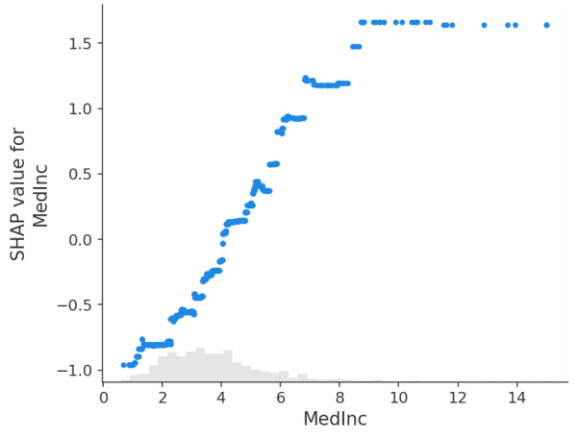
- Badamy, jak zmienia się predykcja modelu dla danej próbki, jeśli do określonego podzbioru cech dołożymy dodatkową cechę *p*.
- Wpływ cechy *p* badamy dla wszystkich możliwych podzbiorów.
- Sumujemy wpływy, uśredniamy wynik.
- Możemy w ten sposób uzyskać wyjaśnienia konkretnych przykładów albo całego modelu - uśredniając wartości na całym zbiorze danych.
- W praktyce: nie trenujemy dodatkowych modeli, stosujemy heurystyki, dedykowane algorytmy dla określonych typów modeli (np. dla drzew), sprytne maskowanie wartości cech.

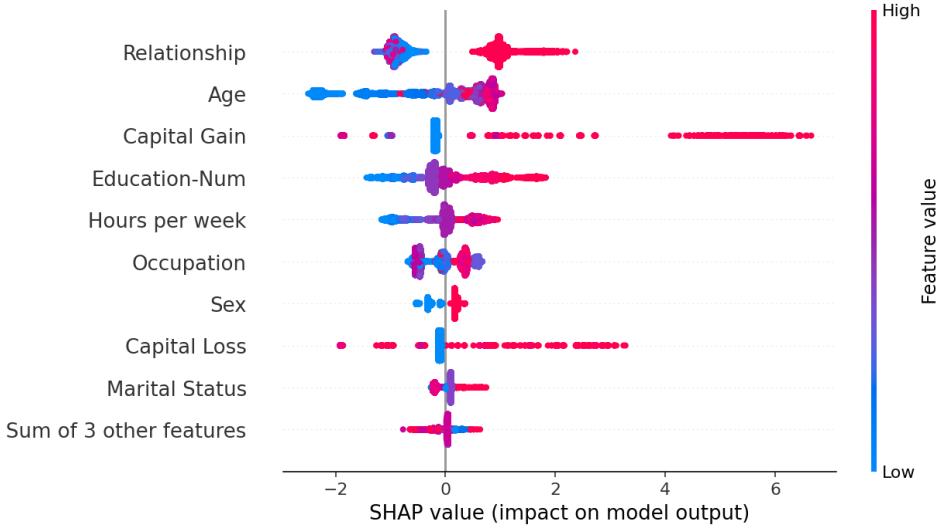
Przykład - Titanic

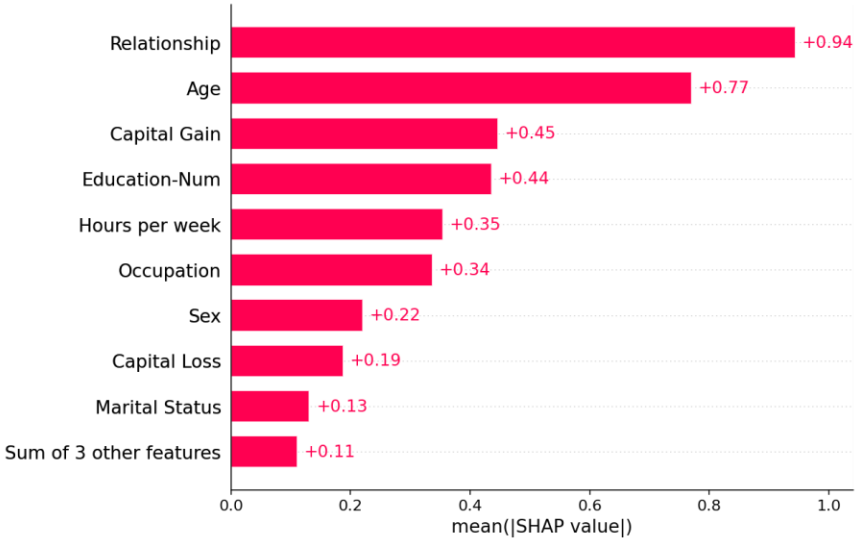


Przykładowe wykresy z wykorzystaniem SHAP









Materiały

- https://ema.drwhy.ai/LIME.html
- https://ema.drwhy.ai/shapley.html
- https://shap.readthedocs.io/en/latest/