Laboratorium 2 - Interpretowalność modeli ML

Mateusz Łopaciński

26 października 2025

1 Wprowadzenie

W ramach laboratorium przeprowadzono analizę interpretowalności modeli ML z wykorzystaniem metody LIME. Eksperymenty obejmowały analizę danych tabularnych (zbiór Adult) oraz obrazów z wykorzystaniem różnych sieci neuronowych.

2 Analiza tabularna - LIME

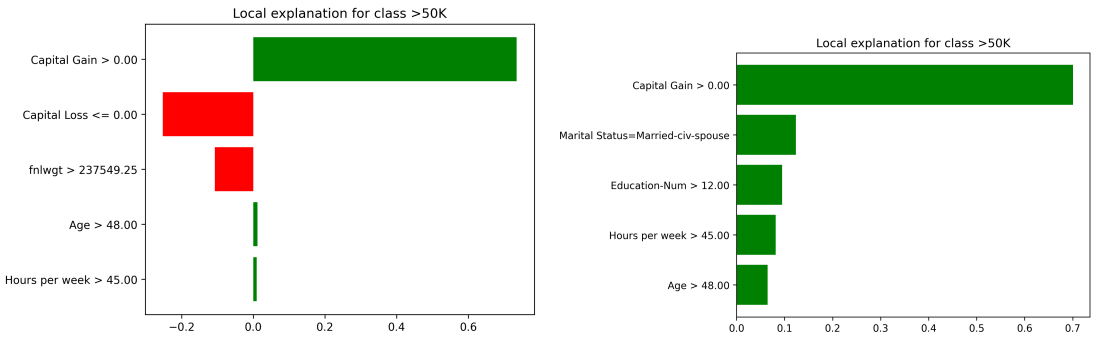
2.1 Porównanie modeli

Wykorzystano MLPClassifier (3 warstwy: 200-100-50 neuronów, ReLU, Adam) zamiast XGBoost z tutorialu. MLP osiągnął dokładność 81.48%.

2.2 Analiza trzech przykładów z tutorialu

2.2.1 Przykład 1653

Oba modele poprawnie przewidziały >50K. XGBoost silnie podkreślał Capital Gain jako główny czynnik, podczas gdy MLP rozłożył wagę między Capital Gain, Capital Loss (ujemny) i fnlwgt (ujemny).

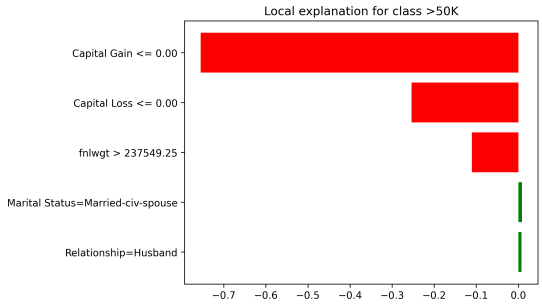


(a) MLPClassifier (b) XGBoost

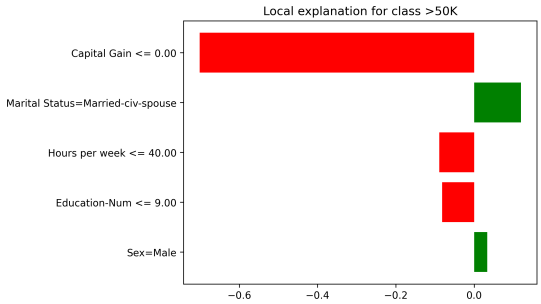
Rysunek 1: Wyjaśnienia LIME dla przykładu 1653

2.2.2 Przykład 92

Oba modele poprawnie przewidziały <=50K. Oba interpretowały brak Capital Gain jako silny negatywny czynnik, ale różniły się wagi dla Marital Status i Hours per week.



(a) MLPClassifier

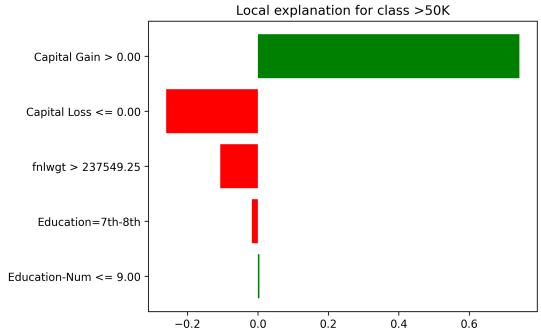


(b) XGBoost

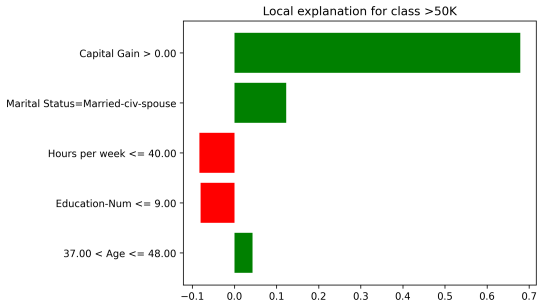
Rysunek 2: Wyjaśnienia LIME dla przykładu 92

2.2.3 Przykład 18

Oba modele poprawnie przewidziały <=50K. Oba interpretowały Capital Gain jako pozytywny czynnik, ale MLP przypisał niższą wagę niż XGBoost. Czynnik Hours per week był negatywny dla obu modeli.



(a) MLPClassifier



(b) XGBoost

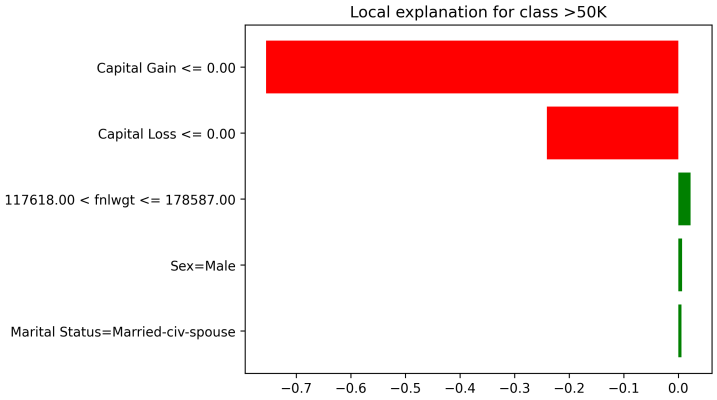
Rysunek 3: Wyjaśnienia LIME dla przykładu 18

2.3 Analiza błędnej predykcji MLP

Znaleziono przykład z błędną klasyfikacją MLP (etykieta: >50K, predykcja: <=50K, 85.1% pewności). LIME ujawnił główne czynniki zmyłające model:

- Capital Gain <= 0.00 (waga: -0.726) - brak zysków kapitałowych jako silny negatywny czynnik
- Capital Loss <= 0.00 (waga: -0.253) - brak strat kapitałowych jako negatywny czynnik
- fnlwgt w zakresie 117618-178587 (waga: +0.019) - minimalny pozytywny wpływ
- Sex=Male (waga: +0.007) - bardzo niska waga pozytywna
- Marital Status=Married-civ-spouse (waga: +0.007) - bardzo niska waga pozytywna

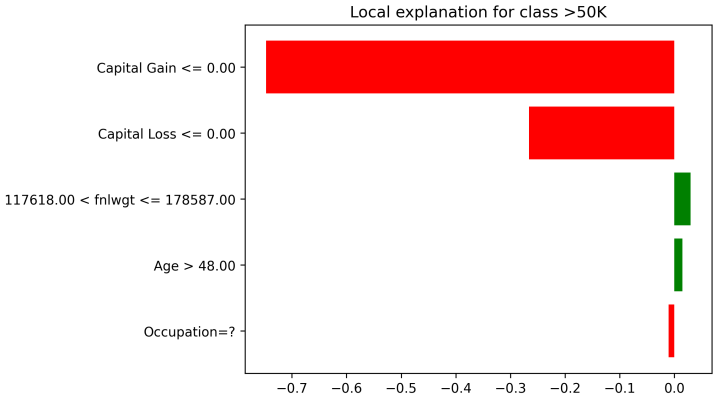
Model został zmyłony przez brak wskaźników kapitałowych, które przeważyły nad innymi czynnikami dochodowymi. LIME pokazuje, że model lokalnie przewiduje 15.4% prawdopodobieństwa dla >50K, ale globalna predykcja MLP wynosi 14.9%.



Rysunek 4: Wyjaśnienia LIME dla błędnej predykcji MLP

2.4 Interesujący przypadek

Przeanalizowano osobę z wysokim wykształceniem (Education-Num $>$ 12) ale niskim dochodem ($<$=50K). Ten przypadek jest interesujący, ponieważ wykształcenie powinno korelować z dochodem, ale tutaj osoba z wysokim wykształceniem ma niski dochód. MLP poprawnie przewidział $<$=50K, pokazując, że inne czynniki (wiek, zawód, Capital Gain/Loss) przeważyły nad wykształceniem. LIME ujawnił, że model rozpoznał inne czynniki jako ważniejsze niż samo wykształcenie.



Rysunek 5: Wyjaśnienia LIME dla interesującego przypadku

3 Analiza obrazów - LIME

3.1 Wybór obrazu

Wybrano obraz smartfona z klawiaturą (smartphone.jpg) spoza zbioru ImageNet. Obraz przedstawia telefon komórkowy leżący obok fragmentu klawiatury komputerowej na drewnianym blacie.

Rysunek 6: Oryginalny obraz smartfona z klawiaturą

3.2 Analiza z wykorzystaniem różnych sieci

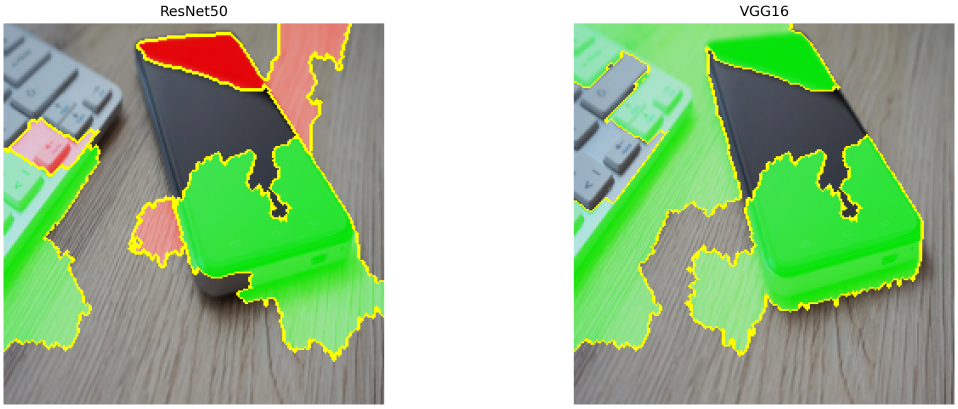
Wykorzystano dwie sieci neuronowe: ResNet50 i VGG16.

ResNet50:

- Najbardziej prawdopodobna klasa: cellular_telephone (57.8%) - telefon komórkowy
- Pozostałe klasy: iPod (20.7%), notebook (9.8%), remote_control (4.1%)

VGG16:

- Najbardziej prawdopodobna klasa: iPod (75.8%) - odtwarzacz iPod
- Pozostałe klasy: cellular_telephone (13.1%), notebook (5.8%), space_bar (1.1%)



Rysunek 7: Porównanie wyjaśnień LIME dla ResNet50 i VGG16

3.3 Analiza po usunięciu kluczowych cech

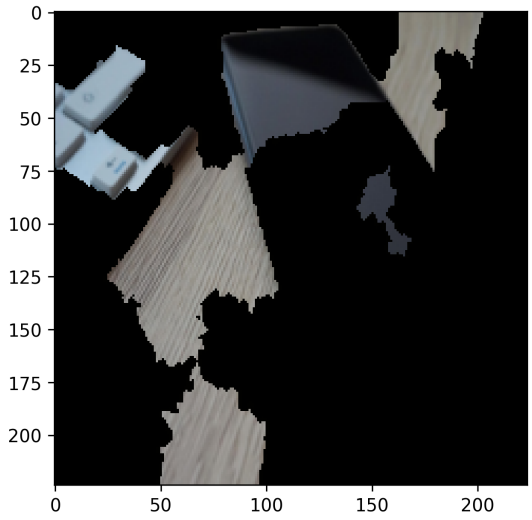
Po usunięciu najważniejszych superpikseli (negative_only=True):

ResNet50:

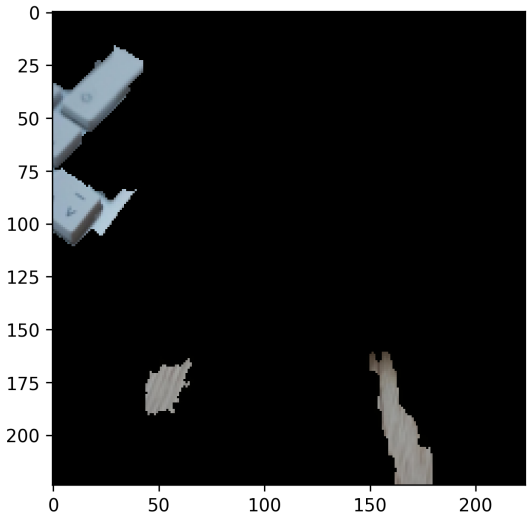
- Nowa najbardziej prawdopodobna klasa: pencil_sharpener (96.0%) - ostrzałka

VGG16:

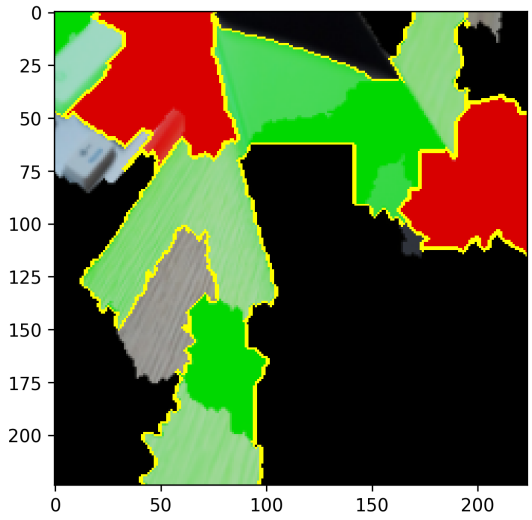
- Nowa najbardziej prawdopodobna klasa: space_shuttle (33.2%) - statek kosmiczny
- Druga najbardziej prawdopodobna klasa: syringe (13.8%) - strzykawka



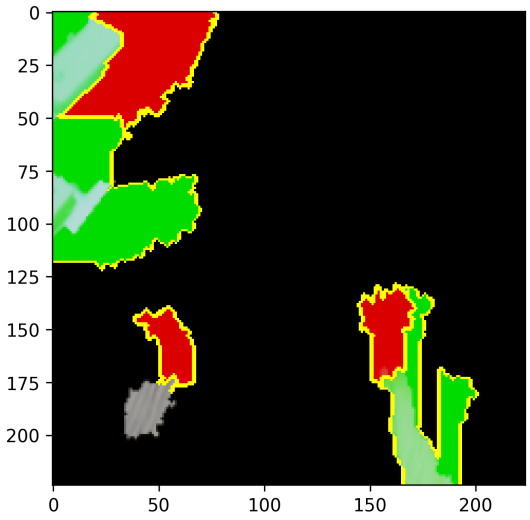
(a) ResNet50 - bez kluczowych cech



(b) VGG16 - bez kluczowych cech



(a) ResNet50 - z wyjaśnieniem



(b) VGG16 - z wyjaśnieniem

4 Wnioski

LIME skutecznie identyfikuje kluczowe cechy wpływające na predykcje (Capital Gain/Loss, wykształcenie, wiek), ale jest to tylko lokalna aproksymacja modelu. LIME może popełniać błędy i nie zawsze dokładnie odzwierciedla globalne zachowanie modelu - pokazuje przybliżoną ważność cech wybranych przez rzeczywisty model. Porównanie MLP z XGBoost wykazało różne wzorce interpretowalności, co pokazuje, że LIME ujawnia różnice w sposobie uczenia się modeli.

W analizie obrazów, LIME skutecznie identyfikował kluczowe regiony, ale różne sieci (ResNet50 vs VGG16) interpretowały te same regiony inaczej. Usunięcie kluczowych cech prowadziło do drastycznej zmiany klasyfikacji.