

Tehnike razlaganja v strojnem ucenju

Explainable Artificial Intelligence (XAI): LIME, SHAP in Grad-CAM

Študent: Dušan Todorović 89232094

Mentor: dr. Branko Kavšek

Zakaj je razlaganje modelov pomembno?

Motivacija in pomen razlozljivosti v praksi

Medicina

Modeli za diagnostiko morajo biti razumljivi zdravnikom — visoka točnost sama po sebi ni dovoljna za varno klinično odločitveno podporo.

Finance

Preglednost pri ocenjevanju kreditnega tveganja in avtomatiziranih finančnih odločitvah je zahteva regulatorjev.

Etika in pravo

Uredba GDPR zagotavlja posamezniku pravico do razlage avtomatiziranih odločitev, ki nanj vplivajo.

Zaznavanje napak

Razlozljivost razkriva pristranskosti in sistemske napake, ki so v kompleksnih modelih sicer skrite.

Temeljni pojmi razlozljivosti

Razlozljivost (Explainability)

Uporaba dodatnih metod za pojasnitev vedenja kompleksnih modelov, ki niso sami po sebi razumljivi.

Lokalna interpretacija

Pojasnjuje, zakaj je model sprejel konkretno odločitev za posamezno instanco.

Model-agnosticne metode

Metode, ki se lahko uporabijo za katerikoli model — LIME in SHAP obravnavata model kot black box.

Interpretabilnost

Sposobnost modela, da je neposredno razumljiv cloveku (npr. linearna regresija, odlocitveno drevo).

Globalna interpretacija

Opisuje splosno vedenje modela prek celotnega podatkovnega niza.

Post-hoc metode

Metode, ki se uporabijo po treniranju modela, brez poseganja v njegovo notranjo strukturo.

LIME — Lokalna razlaga modela

Local Interpretable Model-agnostic Explanations | Ribero, Singh & Gustrini (2018)

OSNOVNA IDEJA

Lokalno aproksimira kompleksen model z enostavnim linearnim modelom v okolici opazovane instance. Model obravnava kot black box.

POSTOPEK

Generiranje perturbacij v blizini instance

Napovedi originalnega modela za perturbacije

Utezevanje vzorcev glede na oddaljenost

Učenje lokalnega linearnega modela

Koeficienti = razlaga prispevkov atributov

PREDNOSTI

- + Deluje z vsakim modelom (model-agnosticno)
- + Intuitivna in hitro razumljiva razlaga
- + Hitra izvedba v praksi

OMEJITVE

- Nestabilnost rezultatov med ponovitvami
- Obcutljivost na izbiro parametrov (sirina jedra)
- Brez globalnega vpogleda v model

SHAP — Razlaga s Shapleyevimi vrednostmi

SHapley Additive exPlanations | Lundberg & Lee (2017)

OSNOVNA IDEJA

Metoda temelji na Shapleyevih vrednostih iz kooperativne teorije iger. Vsak atribut je 'igralac', ki prispeva h koncnemu rezultatu. Zagotavlja konsistentnost in lokalno tocnost.

VARIANTE METODE

TreeSHAP

Optimizirano za drevesne modele (Random Forest, XGBoost)

KernelSHAP

Model-agnosticna razllica — deluje z vsakim modelom

DeepSHAP

Prilagojeno za globoke nevronske mreze

Aditivni model: $f(x) = \phi_0 + \sum_{i=1}^M \phi_i$

PREDNOSTI

- + Teoreticno utemeljena metoda (teorija iger)
- + Konsistentna in stabilna razlaga
- + Lokalna in globalna interpretacija
- + Bogata vizualizacija (bar, beeswarm, waterfall)

OMEJITVE

- Racunsko zahtevnejse od LIME
- Zahteva vecjo tehnicno znanje za interpretacijo
- Visoki racunski stroški za vecje modele

Grad-CAM — Vizualna razlaga nevronskih mrež

Gradient-weighted Class Activation Mapping | Selvaraju et al. (2017)

OSNOVNA IDEJA

Razvita za konvolucijske nevronske mreže (CNN). Gradienti izhoda glede na aktivacije zadnje konvolucijske plasti se uporabijo za generiranje toplotne karte (heatmap), ki pokaže kje je model 'gledal'.

POSTOPEK

CNN klasificira vhodno sliko

Izračun gradientov za ciljni razred

Prostorsko povprečenje gradientov po kanalih

Utežena kombinacija aktivacijskih kart

ReLU + projekcija heatmap na originalno sliko

PREDNOSTI

- + Intuitivna vizualna razlaga (heatmap)
- + Ne zahteva sprememb arhitekture modela
- + Učinkova za medicino in računalniški vid

OMEJITVE

- Omejena na CNN modele
- Manj natančna kot metode na ravni pikslov
- Ne ponuja kolicinskih prispevkov kot SHAP

Primerjava metod: LIME, SHAP in Grad-CAM

Pregled ključnih razlik po kategorijah

Znacilnost	LIME	SHAP	Grad-CAM
Vrsta metode	Model-agnosticna	Model-agnosticna	Model-specificna (CNN)
Raven interpretacije	Lokalna	Lokalna + globalna	Lokalna (vizualna)
Teoreticna osnova	Lok. aproksimacija	Teorija iger (Shapley)	Gradienti v CNN
Stabilnost	Nizja	Visoka	Srednja
Vrsta podatkov	Tabele, besedilo, slike	Tabele, besedilo, slike	Slike
Vizualizacija	Seznam atributov	Bar, beeswarm, waterfall	Toplotna karta
Racunska zahtevnost	Nizka	Visoka	Srednja

Eksperiment: nabor podatkov Titanic

Metodologija, obdelava podatkov in izbira modela

OPIS IN PRIPRAVA

Nabor podatkov OpenML Titanic. Binarna klasifikacija: ali je potnik preživel? Vhodne spremenljivke: spol, starost, razred, cena vozovnice, kraj vkrcanja.

KORAKI PRIPRAVE

- Odstranjeni stolpci: boat, body, name, ticket

(preprečevanje data leakage efekta)

- Delitev: 80 % učni / 20 % testni niz (stratifikacija)
- Stevil. spr.: imputacija z mediano + StandardScaler
- Kat. spr.: imputacija z modusom + one-hot encoding
- Model: RandomForestClassifier (500 dreves)

REZULTATI EVALVACIJE

0.809

Accuracy

0.727

Precision

0.800

Recall

0.762

F1-mera

0.879

ROC AUC

0.849

Avg. Precision

Test data leakage: accuracy 0.943, AUC 0.983

Napihnjene vrednosti — metodolosko neustrezno, model uporablja informacije o izidu.

Interpretacija rezultatov: LIME in SHAP

Lokalna razlaga posameznih napovedi in globalna analiza modela

Ključni atributi (SHAP globalna analiza):

spol (zenska)

spol (moski)

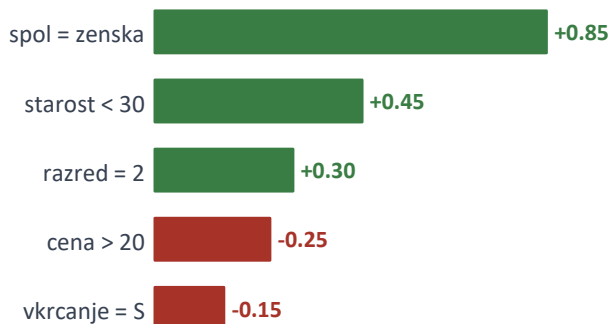
cena vozovnice

razred

starost

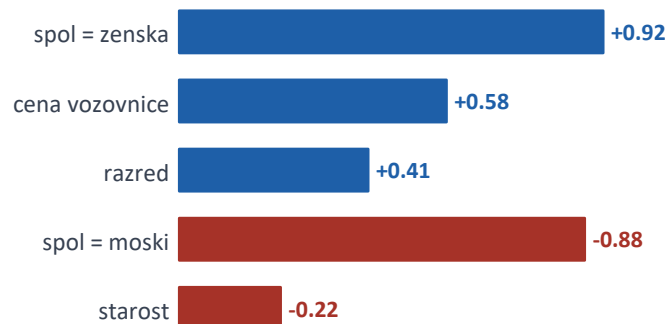
LIME — LOKALNA RAZLAGA

Primer: potnik, $P(\text{prezivel}) = 0.61$



SHAP — GLOBALNA IN LOKALNA RAZLAGA

Konsistentna razlaga — Shapleyeve vrednosti



Oba modela potrjujeta: spol, razred in cena vozovnice so ključni dejavniki prezivljanja — skladno z zgodovinskimi podatki.

ZAKLJUČEK

Pregled ključnih ugotovitev

01

Razlozljivost je strateski imperativ

Interpretabilnost postaja nujna zahteva v medicini, financah in pri eticni uporabi UI — ne zgolj tehnicni dodatek.

02

LIME — hitro in intuitivno

Lokalna model-agnostična razlaga z aproksimacijo. Primerna za hitre vpoglede, a manj stabilna med ponovitvami.

03

SHAP — zanesljivo in celovito

Teoretično utemeljena metoda na osnovi Shapleyevih vrednosti. Omogoča tako lokalno kot globalno analizo s konsistentnimi rezultati.

04

Grad-CAM — vizualna razlaga CNN

Toplotna karta območij slike z največjim vplivom. Nepogrešljivo orodje pri računalniškem vidu in medicinski diagnostiki.

HVALA NA POZORNOST

Literatura:

- *Molnar, C. (2022). Interpretable Machine Learning. <https://christophm.github.io/interpretable-ml-book/>*
- *Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the AAAI Conference on Artificial Intelligence.*
- *Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS).*
- *Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision.*