UP FAMNIT | RAZISKOVALNI SEMINAR | 2026

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 tocnost sama po sebi ni dovoljna za varno klinicno odlocitveno podporo.

Finance

Preglednost pri ocenjevanju kreditnega tveganja in avtomatiziranih financnih odlocitvah je zahteva regulatorjev.

Etika in pravo

Uredba GDPR zagotavlja posamezniku pravico do razlage avtomatiziranih odlocitev, 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.

Interpretabilnost

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

Lokalna interpretacija

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

Globalna interpretacija

Opisuje splosno vedenje modela prek celotnega podatkovnega niza.

Model-agnosticne metode

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

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 | Riberio, 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

Ucenje 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 'igralec', ki prispeva h koncnemu rezultatu. Zagotavlja konsistentnost in lokalno tocnost.

VARIANTE METODE

TreeSHAP	Optimizirano za drevesne modele (Random Forest, XGBoost)
KernelSHAP	Model-agnosticna razlicica — 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 tehnično znanje za interpretacijo

- Visoki racunski stroski za vecje modele

Grad-CAM — Vizualna razlaga nevronskih mrez

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

OSNOVNA IDEJA

Razvita za konvolucijske nevronske mreze (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

Izracun gradientov za ciljni razred

Prostorsko povprecenje gradientov po kanalih

Utezena kombinacija aktivacijskih kart

ReLU + projekcija heatmap na originalno sliko

PREDNOSTI

+ Intuitivna vizualna razlaga (heatmap)

+ Ne zahteva sprememb arhitekture modela

+ Ucinkova za medicino in racunalniski vid

OMEJITVE

- Omejena na CNN modele

- Manj natancna kot metode na ravni pikslov

- Ne ponuja kolicinskih prispevkov kot SHAP

Primerjava metod: LIME, SHAP in Grad-CAM

Pregled kljucnih 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 1: nabor podatkov Titanic

Metodologija, obdelava podatkov in izbira modela

OPIS IN PRIPRAVA

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

KORAKI PRIPRAVE

- Odstranjeni stolpci: boat, body, name, ticket

  (preprecevanje data leakage efekta)

- Delitev: 80 % ucni / 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

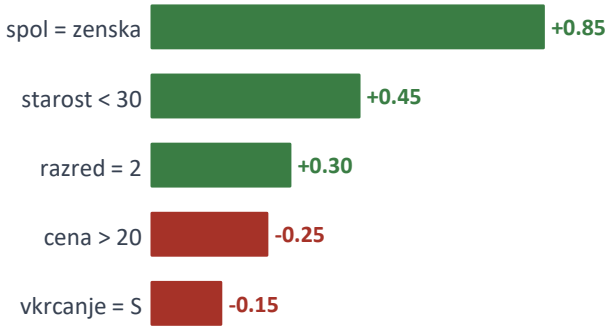
Lokalna razlaga posameznih napovedi in globalna analiza modela

Kljucni atributi (SHAP globalna analiza):

spol (zenska) | spol (moski) | cena vozovnice | razred | starost

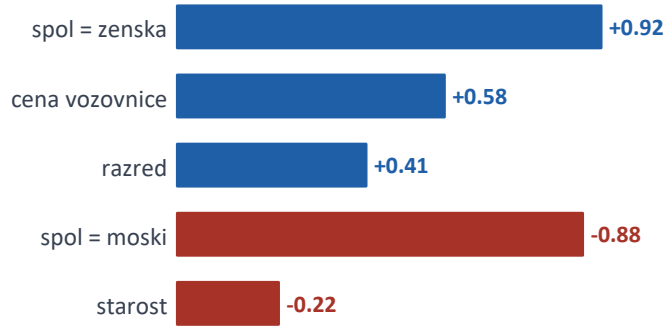
LIME — LOKALNA RAZLAGA

Primer: potnik, P(prezivel) = 0.61



SHAP — GLOBALNA IN LOKALNA RAZLAGA

Konsistentna razlaga — Shapleyeve vrednosti



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

Eksperiment: nabor podatkov Fashion-MNIST

CNN model, vizualna interpretacija in primerjava metod

OPIS IN PRIPRAVA

Podatkovni nabor Fashion-MNIST: 10 razredov oblacil, 70.000 sivinskih slik (28×28 px). Klasifikacija z CNN modelom, implementiranim v PyTorch.

Razlaga barv pri prikazih Grad-CAM in SHAP

Grad-CAM: tople barve kažejo najpomembnejše regije slike. Hladne barve pomenijo manj pomembne regije. Prikazuje le pozitivno podporo ciljnemu razredu.

Pri SHAP rdeča barva označuje pozitiven prispevek k napovedi, modra pa dele slike, ki zmanjšujejo verjetnost opazovanega razreda.

REZULTATI EVALVACIJE

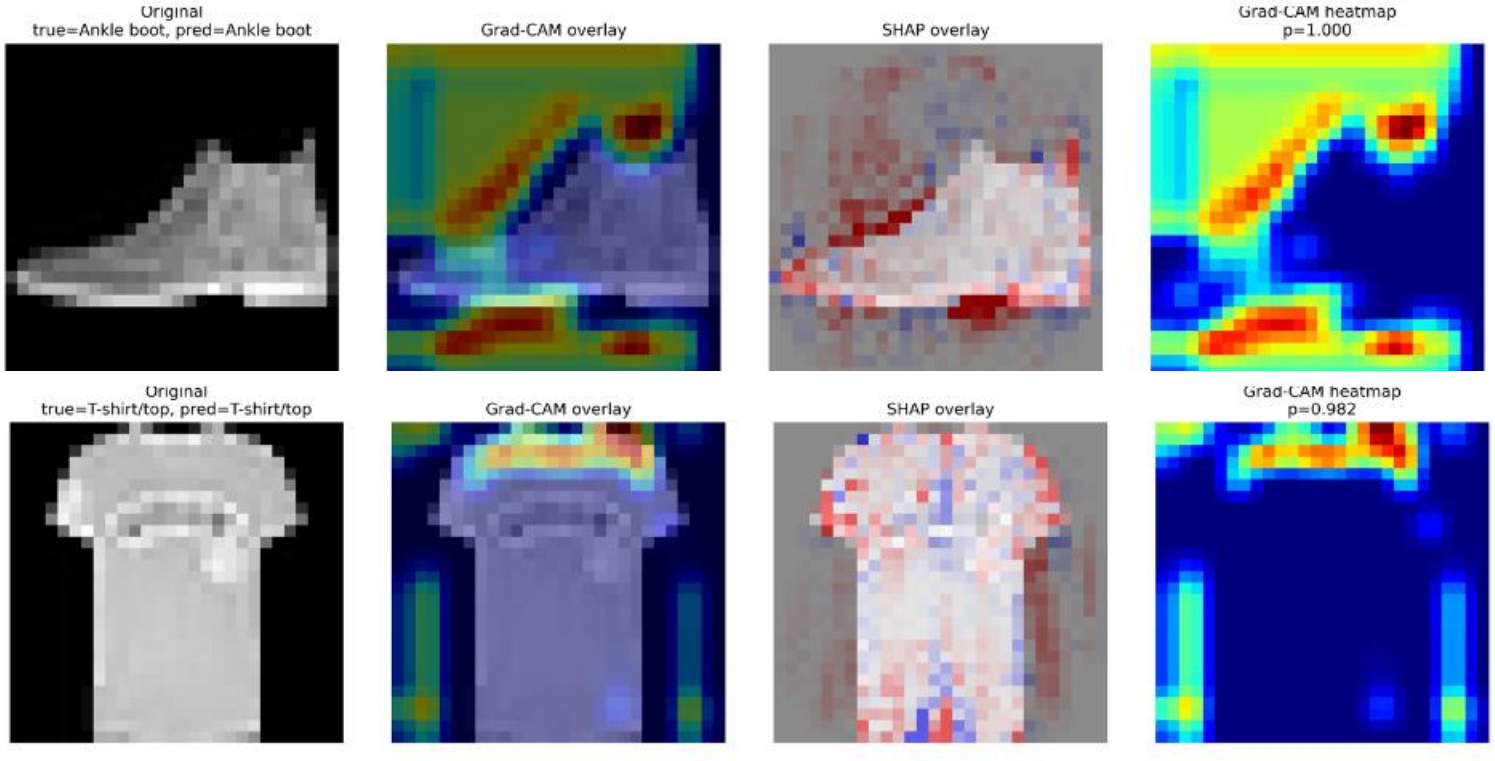
0.914 Accuracy	**0.914** Precision
0.914 Recall	**0.913** F1-mera
0.914 Macro F1	**0.913** Macro Recall

Grad-CAM: vizualizacija regij slike | SHAP: kolicinski prispevki pikslov

Komplementarni metodi: Grad-CAM pokaze kjer, SHAP pokaze koliko posamezni deli prispevajo.

Interpretacija rezultatov: Grad-CAM in SHAP

Vizualna razlaga CNN odlocitev na slikah Fashion-MNIST



Grad-CAM pokaze kje model gleda, SHAP pa kolicinsko loci pozitivne in negativne prispevke — komplementarni metodi za globlje razumevanje CNN.

ZAKLJUCEK

Pregled kljucnih 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-agnosticna razlaga z aproksimacijo. Primerna za hitre vpoglede, a manj stabilna med ponovitvami.

03

SHAP — zanesljivo in celovito

Teoreticno utemeljena metoda na osnovi Shapleyevih vrednosti. Omogoca tako lokalno kot globalno analizo s konsistentnimi rezultati.

04

Grad-CAM — vizualna razlaga CNN

Toplotna karta obmocij slike z najvecjim vplivom. Nepogrešljivo orodje pri racunalniskem 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.*

- *Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv preprint arXiv:1708.07747.*
- *Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Sys-tems.*
- *Vanschoren, J., van Rijn, J. N., Bischl, B., & Torgo, L. (2013). OpenML: Networked Science in Machine Learning. SIGKDD Explorations, 15(2), 49–60.*
- *Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.*
- *Neptune.ai. (2023). Explainability, Auditability and Interpretability in Machine Learning. Dostopno na: https://neptune.ai/blog/explainability-auditability-ml-definitions-techniques-tools*
- *Analytics Vidhya. (2021). Model Explainability in Machine Learning. Dostopno na: https://www.analyticsvidhya.com/blog/2021/11/model-explainability/*
- *DataCamp. (2022). Explainable AI: Understanding and Trusting Machine Learning Models. Dostopno na: https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models*
- *Kaggle. Machine Learning Explainability Course. Dostopno na: https://www.kaggle.com/learn/machine-learning-explainability*