

Rețele neuronale și deep learning

Cursul 11: Inteligența artificială explicabilă: SHAP, LIME, metode bazate pe gradient și stabilitatea explicațiilor

Prof. dr. Daniel Traian PELE

Academia de Studii Economice din București (ASE), Departamentul de Statistică și Econometrie
Academia Română, Institutul de Prognoză Economică (IPE)
IDA Institute Digital Assets ▪ AI4EFin ▪ MSCA Digital Finance

Obiectivele cursului

- Să explicați de ce băncile, asigurătorii și autoritățile de reglementare cer explicații pentru o rețea neuronală și să clasificați metodele inteligenței artificiale explicabile (XAI).
- Să calculați explicațiile aditive Shapley (SHAP) și explicațiile locale interpretabile, independente de model (LIME), cu formulele complete, și să le interpretați pentru un client și pentru un portofoliu.
- Să folosiți metodele bazate pe gradient (saliența, gradienții integrați, SmoothGrad) și să verificați axiomele lor pe un model real de detectare a fraudei.
- Să testați dacă explicațiile sînt concordante, fidele modelului și stabile la schimbarea valorii inițiale, a eșantionului și a perioadei.
- Să formulați o explicație contrafactuală pe care un solicitant respins o poate pune în practică.

Abrevieri

AI	inteligență artificială
AOPC	aria de deasupra curbei de perturbare
ASE	Academia de Studii Economice din București
AUC	aria de sub curba ROC
EBA	Autoritatea Bancară Europeană
EU	Uniunea Europeană (abrevierea engleză, în citări)
EUR	euro
GDPR	Regulamentul general privind protecția datelor
GNN	rețea neuronală pe grafuri
HMEQ	setul de date privind creditele garantate cu locuința
ICE	așteptarea condiționată individuală
IG	gradienți integrați
IPE	Institutul de Prognoză Economică, Academia Română
JS	divergența Jensen-Shannon
KL	divergența Kullback-Leibler
LIME	explicații locale interpretabile, independente de model
MAD	abaterea absolută mediană
MLP	perceptron multistrat
PCA	analiza componentelor principale
PD	probabilitatea de nerambursare
PDP	graficul dependenței parțiale
PICBE	Proceedings of the International Conference on Business Excellence
PR-AUC	aria de sub curba precizie-sensibilitate
ReLU	unitate liniară rectificată
ROC	caracteristica de operare a receptorului
SHAP	explicații aditive Shapley
SR	ghidul de supraveghere privind riscul de model (Federal Reserve SR 11-7)
TCAV	testarea cu vectori de activare a conceptelor
UE	Uniunea Europeană
ULB	Université Libre de Bruxelles
USD	dolar american
XAI	inteligență artificială explicabilă
XGBoost	extreme gradient boosting (arbori amplificați prin gradient)

Datele reale ale cursului

Studiul de caz	Date	Modele și împărțire
Credit	Setul de date privind creditele garantate cu locuința (HMEQ), 5 960 de credite, 19,9 % neplăți (Baesens et al., 2016)	perceptron multistrat (MLP), regresie logistică, extreme gradient boosting (XGBoost); 60/20/20 stratificat
Fraudă	Tranzacții cu carduri din setul Université Libre de Bruxelles (ULB), două zile din septembrie 2013; 281 + 211 fraude în eșantion (Dal Pozzolo et al., 2015)	MLP și regresie logistică; antrenare pe ziua 1, test pe ziua 2
Locuințe	Locuințele din California, 20 640 de grupuri de locuințe din recensământul din 1990	MLP de regresie (dependență parțială)

- Fiecare cifră este calculată de analysis_L11.py
  - Valorile inițiale încep de la 20261108
  - Explicațiile sînt calculate pe un procesor Apple M5 Max.
- Variabilele V1–V28 de fraudă sînt componente ale analizei componentelor principale (PCA) publicate de proprietarul datelor: explicațiile numesc componente, nu cîmpuri brute.

Cuprins

- Motivele explicării rețelelor neuronale
- Valorile Shapley și SHAP
- LIME, contrafactuale și efecte globale
- Metode bazate pe gradient pentru rețele neuronale
- Concordanța, fidelitatea și stabilitatea explicațiilor
- Explicabilitatea în guvernanța modelelor

Reglementările cer explicarea deciziilor automate

- **Regulamentul Uniunii Europene (UE) privind inteligența artificială (AI)** (Regulamentul 2024/1689): evaluarea bonității persoanelor fizice este cu risc ridicat (anexa III)
  - Transparență față de utilizatori (art. 13) și dreptul la explicarea deciziilor individuale (art. 86).
- **Regulamentul general privind protecția datelor (GDPR)**, art. 22 și considerentul 71: garanții față de deciziile bazate exclusiv pe prelucrarea automată.
- **Autoritatea Bancară Europeană (EBA)**, ghidul privind acordarea și monitorizarea creditelor (2020): modelele folosite la decizia de creditare trebuie să fie explicabile.
- **Riscul de model**: ghidul de supraveghere al Federal Reserve privind riscul de model (SR 11-7) cere fundamentare conceptuală, validare independentă și monitorizare continuă.
- **Motive de afaceri**: depistarea scurgerilor de informație, încrederea analiștilor de credit, explicarea unei respingeri pentru client.
- **Totuși**: o explicație este un model al modelului și poate fi greșită.

- **În practică:** o explicație care se schimbă fără ca ceva important să se schimbe este un risc de conformitate, nu o garanție.

(Regulation (EU) 2024/1689; SR 11-7)

O clasificare a metodelor de explicare

Metoda	Domeniu	Necesită	În acest curs
Coeficienți, grile de scor	global, intrinsec	un model liniar	reperul logistic
SHAP	local → global	apeluri ale modelului (Kernel); arbori (TreeSHAP)	MLP de credit, XGBoost
LIME	local	apeluri ale modelului	același solicitant
Salientă, gradienți integrați (IG), SmoothGrad	local	gradienți	MLP de fraudă
Graficul dependenței parțiale (PDP), așteptarea condiționată individuală (ICE)	efect global	apeluri ale modelului	MLP pentru locuințe
Contrafactuale	local, acționabil	gradienți sau căutare	solicitantul respins
Importanța prin permutare	global	etichete + apeluri ale modelului	MLP de fraudă

- Metodele post-hoc explică o cutie neagră deja antrenată
  - Modelele interpretabile sînt transparente prin construcție (Rudin, 2019).

(Rudin, 2019)

Studiul de caz pentru credit: compararea a trei modele de scoring

- **Pe scurt:** rețeaua ordonează rău-platnicii mai bine decît grila logistică (aria de sub curba de operare (ROC) (AUC) și aria de sub curba precizie-sensibilitate (PR-AUC) sînt mai mari); o putem explica la fel de bine?

$$\mathbf{h}_1 = \mathrm{ReLU}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1), \quad \mathbf{h}_2 = \mathrm{ReLU}(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2)$$
$$f(\mathbf{x}) = \mathbf{w}_3^\top \mathbf{h}_2 + b_3, \qquad \mathrm{PD}(\mathbf{x}) = \sigma\big(f(\mathbf{x})\big)$$

- **unde**
  - $\mathbf{x}$ = 18 variabile standardizate, cu indicator pentru raportul datorie/venit lipsă (DEBTINC)
  - f = log-șansele de neplată
  - 64 și 32 de unități liniare rectificate (ReLU), dropout 0,1
  - PD = probabilitatea de nerambursare (PD).

Model (test)	AUC	PR-AUC	Brier
Logistic	0,909	0,785	0,080
MLP, 10 valori	0,935	0,842	0,068
XGBoost, 10	0,959	0,889	0,058

- Rata de bază (PR-AUC al unui scor aleator) 0,200.

- **În practică:** la modelul logistic coeficienții sînt explicația
  - Pentru MLP explicația trebuie calculată.

(Baesens, Roesch & Scheule, 2016; calcul propriu)

Exercițiu practic: notebook §1

Cuprins

- Motivele explicării rețelelor neuronale
- Valorile Shapley și SHAP
- LIME, contrafactuale și efecte globale
- Metode bazate pe gradient pentru rețele neuronale
- Concordanța, fidelitatea și stabilitatea explicațiilor
- Explicabilitatea în guvernanța modelelor

Valorile Shapley: o împărțire echitabilă a prognozei între variabile

- **Pe scurt:** fiecare variabilă primește contribuția sa marginală medie pe toate ordinile în care variabilele s-ar putea alătura prognozei.

$$\phi_j = \sum_{S \subseteq \{1,\ldots,d\} \setminus \{j\}} \frac{|S|!\,(d-|S|-1)!}{d!} \Big[v(S \cup \{j\}) - v(S) \Big]$$

- **unde**
  - d = numărul de variabile
  - $v(S)$ = valoarea coaliției S (ce prognozează modelul cînd sînt cunoscute doar variabilele din S)
  - Ponderea este ponderea ordinilor în care j intră imediat după S.

- **Cele patru axiome** (Shapley, 1953) care fac ϕ unic
  - Eficiența: $\sum_j \phi_j = v(\{1..d\}) - v(\emptyset)$
  - Simetria: contribuții egale, cote egale
  - Variabila nulă: o variabilă care nu schimbă niciodată v primește 0
  - Aditivitatea: $\phi(v_1 + v_2) = \phi(v_1) + \phi(v_2)$
- **Cost**: 2^d coaliții
  - Cu 18 variabile, 2^{18} evaluări ale modelului pentru fiecare solicitant: avem nevoie de aproximări.

- **În practică:** eficiența este ceea ce îi trebuie analistului de credit: părțile se adună exact la scor.

(Shapley, 1953)

SHAP: valoarea unei coaliții de variabile

- **Pe scurt:** SHAP definește $v(S) =$ ieșirea așteptată a modelului cînd variabilele din afara lui S sînt extrase dintr-un eșantion de referință.

$$v(S) = \mathbb{E}_{\mathbf{x}_{\bar{S}} \sim \mathcal{B}}\left[f(\mathbf{x}_S, \mathbf{x}_{\bar{S}})\right]$$

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d} \phi_j(\mathbf{x}), \qquad \phi_0 = \mathbb{E}_{\mathcal{B}}[f]$$

- **unde**
  - $\mathcal{B} = 100$ de solicitanți de antrenare (referința)
  - Variabilele din afara lui S sînt înlocuite independent de $\mathbf{x}_S$
  - Explicațiile sînt pe scara log-șanselor, unde efectele se adună.

- **Aditivitatea pe datele noastre**: eroarea maximă $|\phi_0 + \sum_j \phi_j - f(\mathbf{x})| =$ 2,9e-06 pe 400 de solicitanți din test.
- **Model liniar**: $\phi_j = \beta_j(x_j - \mathbb{E}x_j)$ exact (SHAP liniar).
- **Risc**: cu variabile corelate, înlocuirea independentă creează solicitanți nerealiști.

- **În practică:** eșantionul de referință definește "solicitantul mediu"
  - Documentați-l ca pe orice intrare a modelului.

(Lundberg & Lee, 2017)

Exercițiu practic: notebook §2

Kernel SHAP și TreeSHAP: cum se calculează SHAP

- **Pe scurt:** Kernel SHAP estimează o regresie liniară ponderată pe coaliții eșantionate; pentru arbori valorile exacte se obțin în timp polinomial.

$$\min_{\phi} \sum_{\mathbf{z}\in\{0,1\}^d} \pi(\mathbf{z}) \left[v(\mathbf{z}) - \phi_0 - \sum_j \phi_j z_j \right]^2$$

$$\pi(\mathbf{z}) = \frac{d-1}{\binom{d}{|\mathbf{z}|}\,|\mathbf{z}|\,(d-|\mathbf{z}|)}$$

- **unde**
  - $\mathbf{z}$ = indicatorul coaliției
  - Nucleul Shapley π face ca soluția celor mai mici pătrate ponderate să fie egală cu valorile Shapley (Anexa A)
  - Eșantionăm 600 de coaliții pentru fiecare solicitant.

- **Kernel SHAP**: independent de model
  - 16 s pentru 400 de solicitanți.
- **TreeSHAP**: exact pentru ansambluri de arbori (Lundberg et al., 2020).
- **Rețele adînci**: Deep/Gradient SHAP combină valorile Shapley cu gradienții (secțiunea 3).

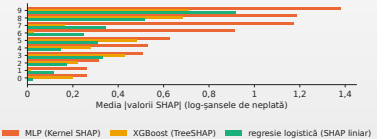
- **În practică:** Kernel SHAP are zgomot de eșantionare: documentați valoarea inițială și numărul de eșantioane.

(Lundberg & Lee, 2017; Lundberg et al., 2020)

Exercițiu practic: notebook §2

Importanța globală: lipsa raportului datorie/venit domină

- **Pe scurt:** valoarea SHAP absolută medie pe variabilă, pentru rețea și cele două repere.



Sursa: Baesens, Roesch și Scheule (2016), datele HMEQ. Date: 400 de solicitanți din test, 100 de solicitanți de referință.

- Pe primul loc la toate trei modelele: DEBTINC_missing
  - La MLP urmează DELINQ și JOB_Other
  - Corelația Spearman a importanțelor MLP–XGBoost 0,55, MLP–logistic 0,79.
- Raportul datorie/venit (DEBTINC) lipsește la 21,3 % dintre credite
  - Rata de neplată este 62,0 % cînd lipsește și 8,6 % cînd este raportat.

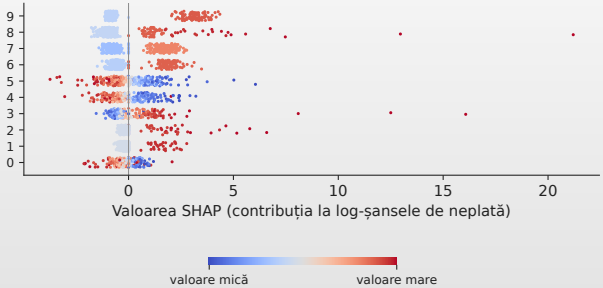
- **În practică:** o explicație poate arăta că modelul notează dosarul, nu debitorul
  - Decideți dacă acest lucru este acceptabil.

(calcul propriu)

NN_L11_shap_global

Diagrama roi: sensul și dispersia fiecărui efect

- **Pe scurt:** fiecare punct este un solicitant; culoarea = rangul valorii variabilei; poziția = valoarea SHAP.



Sursa: Baesens, Roesch și Scheule (2016), datele HMEQ. Date: 400 de solicitanți din test; culoarea = rangul valorii variabilei.

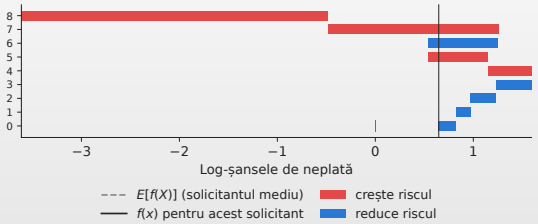
- Liniile de credit cu restanțe (DELINQ) și rapoartele negative (DEROG): valorile mari cresc riscul, așa cum se așteaptă în teoria creditului.

(calcul propriu)

NN_L11_shap_beeswarm

Explicația locală pentru un solicitant respins: diagrama cascadă SHAP

⊡ **Pe scurt:** de la log-șansele medii ϕ_0 la scorul acestui solicitant, variabilă cu variabilă.



Sursa: Baesens, Roesch și Scheule (2016), datele HMEQ. Date: solicitant din test (rîndul 2411); valori în unitățile originale.

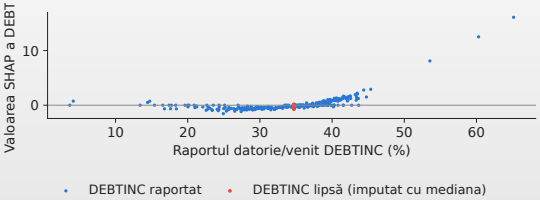
⊡ Solicitant din test (rîndul 2411): PD 65,6 %, log-șanse 0,648 $= -3{,}613 + 4{,}261$
- Cele mai mari contribuții: DEBTINC_missing (+3,13), JOB_Other (+1,74), DELINQ (−0,71).

⊡ **În practică:** notificarea de respingere enumeră principalele motive în această ordine, în cuvinte pe care clientul le înțelege.

(calcul propriu) NN_L11_shap_waterfall

Graficul de dependență: rețeaua a învățat semnalul valorii lipsă

- **Pe scurt:** valoarea SHAP a DEBTINC în funcție de valoarea sa; punctele roșii sînt imputate cu mediana din antrenare.



Sursa: Baesens, Roesch și Scheule (2016), datele HMEQ. Date: 400 de solicitanți din test, MLP.

- Solicitanții cu valoare imputată au aceeași valoare DEBTINC
  - Riscul lor este preluat de indicatorul de lipsă (SHAP mediu $+3{,}13$), nu de DEBTINC.

- **În practică:** alegerea metodei de imputare schimbă explicațiile
  - Păstrați un indicator de lipsă pentru ca motivul să poată fi numit.

(calcul propriu)

NN_L11_shap_dependence

Cuprins

- Motivele explicării rețelelor neuronale
- Valorile Shapley și SHAP
- LIME, contrafactuale și efecte globale
- Metode bazate pe gradient pentru rețele neuronale
- Concordanța, fidelitatea și stabilitatea explicațiilor
- Explicabilitatea în guvernanța modelelor

LIME: un model simplu, fidel în jurul unui punct

- **Pe scurt:** solicitantul este perturbat, perturbările sînt ponderate după apropiere și se estimează un model liniar rar pe răspunsurile rețelei.

$$\xi(\mathbf{x}) = \arg\min_{g \in \mathcal{G}} \sum_{\mathbf{z}} \pi_{\mathbf{x}}(\mathbf{z}) \Big[f(\mathbf{z}) - g(\mathbf{z}) \Big]^2 + \Omega(g)$$

$$\pi_{\mathbf{x}}(\mathbf{z}) = \exp\left(- \frac{\|\mathbf{z} - \mathbf{x}\|^2}{\sigma^2} \right)$$

- **unde**
  - $\mathcal{G}$ = modelele liniare $g(\mathbf{z}) = \beta_0 + \beta^\top \mathbf{z}$
  - Ω = penalizarea complexității
  - $\mathbf{z} \sim \mathcal{N}(\mathbf{x}, \mathbf{I})$ în unități standardizate
  - Lățimea nucleului $\sigma = 0{,}75\sqrt{d} = 3{,}18$ implicit.

- **Pantă locală**, nu o împărțire a prognozei: coeficienții nu se adună la $f(\mathbf{x})$.
- **Trei alegeri** decid rezultatul: vecinătatea, lățimea nucleului, numărul de eșantioane.
- **Solicitantul nostru**: variabila de top la LIME este DELINQ, la SHAP este DEBTINC_missing.

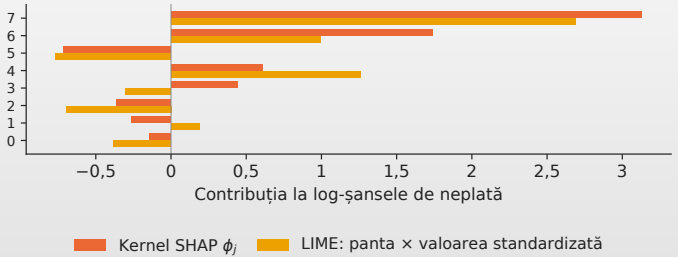
- **În practică:** LIME răspunde la "ce mișcă scorul în jurul acestui client?", iar SHAP la "de ce este acest client departe de medie?".

(Ribeiro, Singh & Guestrin, 2016)

Exercițiu practic: notebook §3

LIME și SHAP pentru același solicitant

- **Pe scurt:** panta LIME înmulțită cu valoarea standardizată, alături de contribuția SHAP, pentru cele mai importante opt variabile SHAP.



Sursa: Baesens, Roesch și Scheule (2016), datele HMEQ. Date: solicitant din test (rîndul 2411); LIME cu 5 000 de eșantioane.

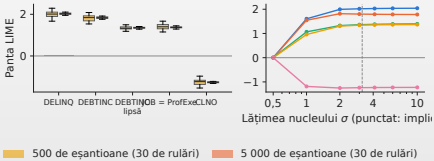
- Corelația Spearman a contribuțiilor absolute: 0,80
  - Un indicator binar este departe de aproximarea sa liniară locală.

(calcul propriu)

NN_L11_lime_vs_shap

Sensibilitatea LIME la propriile setări

- **Pe scurt:** stînga: 30 de rulări cu valori inițiale diferite; dreapta: cele cinci pante cînd se schimbă lățimea nucleului.



Sursa: Baesens, Roesch și Scheule (2016), datele HMEQ. Date: solicitant din test (rîndul 2411); dreapta: aceleași cinci variabile, 5 000 de eșantioane.

- Aceeași variabilă de top: 73 % din rulări cu 500 de eșantioane, 100 % cu 5 000
  - Suprapunerea Jaccard a primelor 5 variabile 0,72, respectiv 1,00
  - Coeficientul de variație 10 %, respectiv 2 %.

- **În practică:** fixați și raportați valoarea inițială, mărimea eșantionului și lățimea nucleului pentru orice explicație LIME care ajunge la un client sau la un auditor.

(calcul propriu)

NN_L11_lime_instability

Explicații contrafactuale: modificările minime care schimbă decizia

- **Pe scurt:** se caută cel mai apropiat solicitant, schimbînd doar variabile acționabile, pe care modelul l-ar accepta.

$$\mathbf{x}' = \arg\min_{\mathbf{x}'} \; \lambda \left(f(\mathbf{x}') - t \right)_+^2 + \sum_{j \in \mathcal{A}} \frac{|x'_j - x_j|}{\mathrm{MAD}_j}$$

- **unde**
  - t = log-șansele-țintă ($-0{,}5$, sub pragul PD de 0,5)
  - $\mathcal{A}$ = variabilele acționabile (LOAN, DEBTINC, NINQ, DELINQ, DEROG, CLNO, MORTDUE)
  - Abaterea absolută mediană (MAD) scalează distanța
  - Valorile brute rămîn ≥ 0, numărătorile sînt rotunjite.

Variabila	acum	necesar
MORTDUE	69 115,0	57 456,5
DEBTINC	34,7	33,4
LOAN	14 500,0	15 700,0

- PD 65,6 % → 39,2 % după aceste schimbări.
- Contrafactualul documentar: raportarea DEBTINC (la mediana 34,7) dă PD 5,1 %.

- **În practică:** "aduceți documentul de venit care lipsește" este un răspuns mai bun pentru client decît orice listă de valori SHAP.

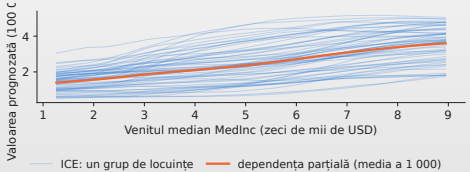
(Wachter, Mittelstadt & Russell, 2017)

Exercițiu practic: notebook §4

Efecte globale: dependența parțială și ICE

- **Pe scurt:** PDP mediază prognoza pe date cînd o variabilă este fixată la o valoare din grilă; ICE păstrează cîte o curbă pentru fiecare observație.

$$\mathrm{PD}_j(z) = \frac{1}{n}\sum_{i=1}^{n} f(z, \mathbf{x}_{i,-j}), \qquad \mathrm{ICE}_{i,j}(z) = f(z, \mathbf{x}_{i,-j})$$



Sursa: StatLib, locuințe din California, recensămîntul din 1990 (Pace și Barry, 1997). Date: MLP; 60 de grupuri de test (ICE), 1 000 pentru PDP.

- MLP pentru locuințele din California (R^2 de test 0,819, liniar 0,620): valoarea prognozată crește de la 1,40 la 3,60 (100 000 USD) pe grila de venit
  - Dispersia curbelor ICE la capăt 0,95.

(Friedman, 2001; Goldstein et al., 2015)

NN_L11_pdp_ice

Cuprins

- Motivele explicării rețelelor neuronale
- Valorile Shapley și SHAP
- LIME, contrafactuale și efecte globale
- Metode bazate pe gradient pentru rețele neuronale
- Concordanța, fidelitatea și stabilitatea explicațiilor
- Explicabilitatea în guvernanța modelelor

Studiul de caz pentru fraudă: o rețea antrenată pe ziua 1 și explicată pe ziua 2

- **Pe scurt:** pentru o rețea gradienții sînt gratuiți: o singură trecere înapoi dă sensibilitatea scorului de fraudă la fiecare intrare.

$$\text{saliency}_j(\mathbf{x}) = \left| \frac{\partial f(\mathbf{x})}{\partial x_j} \right|, \qquad \text{G}\times\text{I}_j(\mathbf{x}) = (x_j - x'_j)\,\frac{\partial f(\mathbf{x})}{\partial x_j}$$

- **unde**
  - f = log-șansele de fraudă ale unui MLP pe V1–V28 și logaritmul sumei
  - $\mathbf{x}'$ = referința = tranzacția legitimă medie din ziua 1
  - Ponderea clasei $\sqrt{\text{dezechilibru}}$, oprire timpurie după PR-AUC ponderat de validare.

- **Test, ziua 2** (ponderi pentru datele complete): PR-AUC MLP 0,815, logistic 0,742
  - AUC 0,961, respectiv 0,964.
- **Saturarea**: într-o zonă plată a lui f gradientul este aproape zero chiar și pentru o variabilă decisivă.
- **Rețelele ReLU**: gradientul este constant pe porțiuni, deci salienta sare între regiunile liniare.

- **În practică:** un analist de fraudă primește lista ordonată a intrărilor care au declanșat alerta, în milisecunde pe tranzacție.

(Simonyan, Vedaldi & Zisserman, 2014; Dal Pozzolo et al., 2015)

Exercițiu practic: notebook §5

Gradienții integrați și completitudinea atribuirilor

- **Pe scurt:** se integrează gradientul pe drumul drept de la referință la intrare; atribuirile se adună la $f(\mathbf{x}) - f(\mathbf{x}')$.

$$\mathrm{IG}_j(\mathbf{x}) = (x_j - x'_j) \int_0^1 \frac{\partial f\left(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}')\right)}{\partial x_j}\, d\alpha$$
$$\approx (x_j - x'_j) \frac{1}{m} \sum_{k=1}^{m} \frac{\partial f\left(\mathbf{x}' + \frac{k-1/2}{m}(\mathbf{x} - \mathbf{x}')\right)}{\partial x_j}$$

- **unde**
  - $m =$ pașii Riemann (regula punctului de mijloc, 64 aici)
  - Completitudinea $\sum_j \mathrm{IG}_j = f(\mathbf{x}) - f(\mathbf{x}')$ rezultă din teorema fundamentală a calculului integral pe drumuri.

- **Axiome** (Sundararajan et al., 2017)
  - Completitudinea; sensibilitatea: o variabilă care schimbă f primește o atribuire nenulă
  - Invarianța la implementare: funcții egale, atribuiri egale
- **Legătura cu Shapley**: IG este valoarea Aumann-Shapley a jocului continuu de-a lungul drumului.
- **Referința contează**: "tranzacția legitimă medie" răspunde la "de ce nu este aceasta o plată normală?".

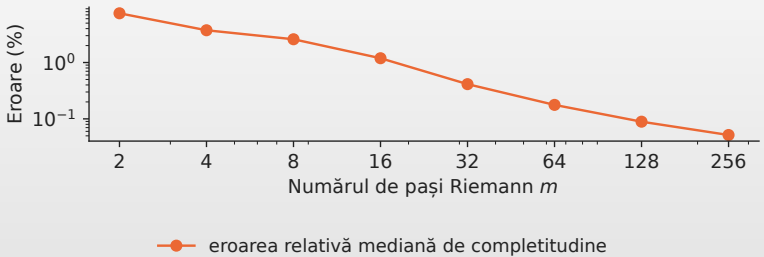
- **În practică:** scrieți referința în documentația modelului: o altă referință înseamnă o altă întrebare.

(Sundararajan, Taly & Yan, 2017)

Exercițiu practic: notebook §5

Completitudinea se verifică atunci cînd integrala este calculată precis

- **Pe scurt:** mediana lui $|\sum_j \mathrm{IG}_j - (f(\mathbf{x}) - f(\mathbf{x}'))|/|f(\mathbf{x}) - f(\mathbf{x}')|$ pe fraudele din ziua 2.



Sursa: ULB Machine Learning Group și Worldline, tranzacții cu carduri (OpenML 1597). Date: 211 fraude din ziua a doua; referința = tranzacția legitimă medie din ziua întîi.

- Eroare de 3,7 % cu 4 pași și de 0,18 % cu 64, pe 211 de fraude
  - Eroarea scade aproximativ ca $1/m^2$, ca la regula punctului de mijloc.

(calcul propriu)

NN_L11_ig_completeness

SmoothGrad și explicațiile la nivel de concept

- **Pe scurt:** se mediază gradienții pe copii perturbate ale intrării pentru a elimina oscilațiile ReLU; sau se testează dacă un concept uman mișcă ieșirea.

$$\mathrm{SG}_j(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n} \frac{\partial f(\mathbf{x}+\varepsilon_k)}{\partial x_j}, \quad \varepsilon_k \sim \mathcal{N}(\mathbf{0}, s^2\mathbf{I})$$

$$S_C(\mathbf{x}) = \nabla h(\mathbf{x})^\top \mathbf{v}_C$$

- **unde**
  - $n = 50$, $s = 0{,}15$ abateri standard
  - Testarea cu vectori de activare a conceptelor (TCAV): $\mathbf{v}_C$ = normala unui clasificator liniar care separă exemplele conceptului C într-un strat ascuns h
  - Scorul TCAV este ponderea intrărilor cu $S_C > 0$.

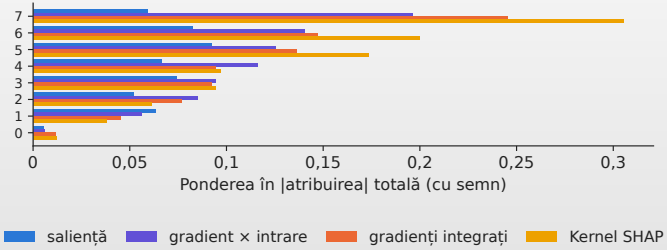
- **SmoothGrad** clarifică hărțile de salienț ale imaginilor și stabilizează gradienții tabelari (secțiunea următoare).
- **TCAV** vorbește limbajul afacerii: “tranzacție de noapte”, “comerciant nou” — dacă asemenea concepte sînt etichetate.
- **Datele noastre**: V1–V28 sînt componente anonimizate, deci aici nu se pot face teste pe concepte.

- **În practică:** explicațiile pe concepte cer concepte etichetate încă de la colectarea datelor.

(Smilkov et al., 2017; Kim et al., 2018)

Patru metode de explicare aplicate unei tranzacții frauduloase

- **Pe scurt:** atribuirile cu semn pentru o fraudă din ziua 2 (scorul median al fraudelor), ca ponderi în atribuirea absolută totală.



Sursa: ULB Machine Learning Group și Worldline, tranzacții cu carduri (OpenML 1597). Date: o fraudă din ziua a doua (scorul median al fraudelor).

- Variabila de top: Kernel SHAP V14, gradienți integrați V14, salienţă V12
  - Scorul de fraudă 99,9 %: salienţa nu ține cont de mărimea abaterii față de referință.

(calcul propriu)

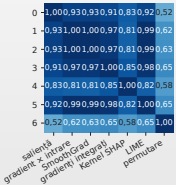
NN_L11_fraud_local

Cuprins

- Motivele explicării rețelelor neuronale
- Valorile Shapley și SHAP
- LIME, contrafactuale și efecte globale
- Metode bazate pe gradient pentru rețele neuronale
- Concordanța, fidelitatea și stabilitatea explicațiilor
- Explicabilitatea în guvernanța modelelor

Concordanța metodelor de explicare

- **Pe scurt:** corelația Spearman a importanțelor globale (media |atribuirii|) pentru șapte metode, pe aceleași fraude.



Sursa: ULB Machine Learning Group și Worldline, tranzacții cu carduri (OpenML 1597). Date: corelația Spearman a importanțelor globale; 211 fraude din ziua a doua.

- IG–SHAP 0,85, salienţă–SHAP 0,83, LIME–SHAP 0,82, permutare–SHAP 0,58
  - Toate metodele pun V12 pe primul loc.

- **În practică:** dezacordul este informație: importanța prin permutare arată de ce are nevoie modelul, atribuirile arată ce folosește.

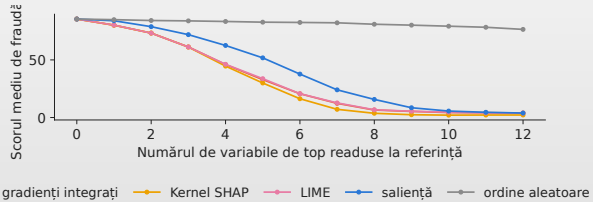
(Breiman, 2001; calcul propriu)

NN_L11_method_agreement

Fidelitatea explicațiilor: testul eliminării variabilelor

- **Pe scurt:** variabilele de top k ale fiecărei fraude sînt readuse la referință; o explicație fidelă face scorul să scadă repede.

$$\mathrm{AOPC} = \frac{1}{K} \sum_{k=1}^{K} \left[f(\mathbf{x}) - f(\mathbf{x}^{(k)}) \right]$$



Sursa: ULB Machine Learning Group și Worldline, tranzacții cu carduri (OpenML 1597). Date: 211 fraude din ziua a doua.

- Aria de deasupra curbei de perturbare (AOPC), în puncte de scor de fraudă: IG 56,2, SHAP 58,4, LIME 56,1, salientă 48,0, ordine aleatoare 3,9.

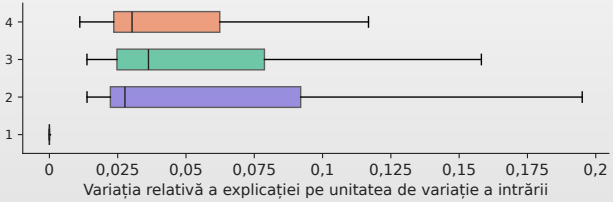
(Samek et al., 2017)

NN_L11_deletion_curve

Robustețea locală a explicațiilor la mici modificări ale intrării

- **Pe scurt:** variația relativă a vectorului de atribuiri pe unitatea unei perturbări mici (raport de tip Lipschitz).

$$\hat{L}(\mathbf{x}) = \frac{\|\mathbf{a}(\mathbf{x}+\varepsilon) - \mathbf{a}(\mathbf{x})\| / \|\mathbf{a}(\mathbf{x})\|}{\|\varepsilon\|}, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, 0.05^2\mathbf{I})$$



Sursa: ULB Machine Learning Group și Worldline, tranzacții cu carduri (OpenML 1597). Date: 60 de fraude din ziua a doua x 10 perturbări (sigma = 0,05 abateri standard).

- Mediane: salientă 0,000, gradient × intrare 0,028, IG 0,030, SmoothGrad 0,036
  - În interiorul unei regiuni liniare a unei rețele ReLU gradientul nu se schimbă.

(Alvarez-Melis & Jaakkola, 2018)

NN_L11_local_robustness

Efectul Rashomon: modele la fel de precise, cu explicații diferite

- **Pe scurt:** rangul fiecărei variabile după media |SHAP| pentru 10 rețele și 10 ansambluri de arbori reantrenate (doar valorile inițiale diferă).



Sursa: Baesens, Roesch și Scheule (2016), datele HMEQ. Date: 10 valori inițiale de antrenare; 200 de solicitanți din test.

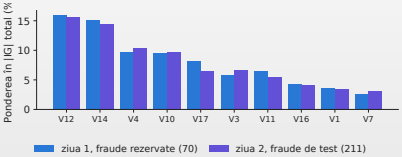
- AUC de test MLP între 0,929 și 0,941
  - Tau Kendall mediu pe perechi (acordul rangurilor importanțelor, 1 = aceeași ordine): MLP 0,83, XGBoost 0,97
  - Aceeași variabilă de top pentru un solicitant în 78 % (MLP) și 83 % (XGBoost) dintre perechile de valori inițiale.

(calcul propriu)

NN_L11_seed_stability

Explicațiile în timp: ziua 1 față de ziua 2

- **Pe scurt:** ponderea în |IG| total pe variabile, pentru fraudele rezervate din ziua de antrenare și pentru fraudele din ziua următoare.



Sursa: ULB Machine Learning Group și Worldline, tranzacții cu carduri (OpenML 1597). Date: model antrenat pe ziua 1; gradienți integrați.

- Spearman 0,98, divergența Jensen-Shannon (JS) (construită pe divergența Kullback-Leibler (KL)) 0,003, 4 variabile comune în primele 5
  - După o zi, explicația este stabilă aici.

- **În practică:** monitorizați lunar distribuția atribuirilor, alături de distribuția scorurilor
  - O deplasare a motivelor poate preceda o scădere a acurateței.

(calcul propriu)

NN_L11_time_shift

Fișa de stabilitate a acestui curs

Sursa schimbării	Măsura	Rezultat
Aleatorismul explicației (LIME)	aceeași variabilă de top în 30 de rulări	73 % (500), 100 % (5 000 de eșantioane)
Valoarea inițială (Rashomon)	tau Kendall al importanțelor globale	MLP 0,83, XGBoost 0,97
Valoarea inițială, un solicitant	aceeași variabilă de top	MLP 78 %, XGBoost 83 %
Metoda	Spearman al importanțelor	0,52 – 1,00
Zgomotul la intrare	variația relativă mediană	IG 0,030, SmoothGrad 0,036
Timpul (ziua 1 $\rightarrow$ ziua 2)	Spearman, divergența JS	0,98; 0,003

⊡ **În practică:** raportați orice explicație împreună cu stabilitatea ei: ce valoare inițială, ce eșantion, ce perioadă și cît se modifică.

(calcul propriu)

Cuprins

- Motivele explicării rețelelor neuronale
- Valorile Shapley și SHAP
- LIME, contrafactuale și efecte globale
- Metode bazate pe gradient pentru rețele neuronale
- Concordanța, fidelitatea și stabilitatea explicațiilor
- Explicabilitatea în guvernanța modelelor

Lista de verificare pentru un audit al explicabilității

- **Scopul**: cine citește explicația (client, analist, validator, autoritate) și ce decizie susține.
- **Setările**: eșantionul de referință, referința, valoarea inițială, numărul de eșantioane, lățimea nucleului — versionate odată cu modelul.
- **Verificările**: aditivitate/completitudine, fidelitate (eliminare), stabilitate (valori inițiale, timp), concordanța a două metode.
- **Echitatea**: atribuirile variabilelor protejate și ale substitutelor lor (vîrstă, sex, cod poștal)
  - O explicație arată utilizarea, nu cauzalitatea.
- **Pentru client**: puține motive, contrafactuale acționabile, fără nume brute de componente (V14 nu înseamnă nimic pentru un client).
- **Alternativa**: dacă un model interpretabil este aproape la fel de precis (aici AUC logistic 0,909 față de MLP 0,935), luați-l în considerare.

În practică: raportul de validare trebuie să conțină explicațiile și testele lor de stabilitate, nu doar acuratețea.

(SR 11-7; Rudin, 2019)

Test 1: adunarea valorilor SHAP

- Pentru solicitantul nostru valoarea de bază este $-3{,}613$, iar valorile SHAP se adună la $4{,}261$.
  - Care sînt log-șansele de neplată date de model și de ce sînteți siguri?

Răspuns

- 0,648: din axioma eficienței, $f(\mathbf{x}) = \phi_0 + \sum_j \phi_j$
  - PD este $\sigma(0{,}648) = 65{,}6\,\%$.

Test 2: un LIME instabil

- Doi analiști explică același solicitant respins cu LIME și raportează motive principale diferite.
  - Este modelul inconsecvent?

Răspuns

- Nu neapărat: LIME eșantionează perturbări
  - Cu 500 de eșantioane variabila de top a coincis în 73 % din rulările noastre
  - Fixați valoarea inițială, creșteți eșantionul, comparați cu SHAP.

Test 3: alegerea referinței IG

- Un coleg folosește vectorul nul ca referință pentru gradienții integrați, pe variabile de fraudă standardizate.
  - La ce întrebare răspunde explicația?

Răspuns

- "De ce diferă această tranzacție de tranzacția medie (inclusiv fraudele)?"
  - Referința noastră, tranzacția legitimă medie, întreabă de ce nu este o plată normală.

Recapitulare

- **SHAP**: valori Shapley cu verificarea exactă a aditivității
  - Lipsa raportului datorie/venit determină rețeaua de credit.
- **LIME**: o pantă locală care depinde de valoarea inițială, de eșantion și de lățimea nucleului
  - Utilă, dar raportați setările.
- **Gradienți**: gradienții integrați satisfac completitudinea (eroare de 0,18 % la 64 de pași) și sînt la fel de fideli ca SHAP, la un cost mult mai mic.
- **Stabilitatea**: reantrenarea cu altă valoare inițială reordonează variabilele (tau Kendall 0,83 pentru MLP)
  - Explicațiile trebuie testate ca modelele.
- **Urmează**: cursul 12, modelele lingvistice mari, unde explicațiile devin text.

Cercetările grupului nostru: XAI în fraudă și piețele de energie

- **Stabilitatea SHAP în detectarea fraudei cu informație din graf**: pe datele de fraudă IEEE-CIS se compară contribuțiile SHAP ale unui model XGBoost cu reprezentări dintr-o rețea neuronală pe grafuri (GNN), sub validare aleatoare și validare înainte în timp, cu un control gaussian de aceeași dimensiune și covariantă
  - Dacă modelul cu graf pare mai stabil depinde de modul în care se măsoară schimbarea, un protocol riguros pentru stabilitatea explicațiilor (Diaconu & Pele, 2026, trimis la Expert Systems with Applications; cod: QuantLet GNN_XAI_Stability).
- **Avertizarea timpurie explicabilă a vîrfurilor de preț** la energia electrică din România (Andrei, Pele, Cramer, Amza & Diaconu, 2026, Proceedings of the International Conference on Business Excellence (PICBE)) și **detectarea anomaliilor de preț cu SHAP** (Andrei & Pele, 2025, Preprints).
- **Explicabilitatea pe mai multe orizonturi** a prognozelor de preț: ce factori contează la fiecare orizont pe piața liberalizată din România (Andrei, Pele & Cramer, 2026, PICBE).
- Rezultatele sînt cele din lucrări; cifrele de pe celelalte slide-uri provin din datele noastre.

(Academia de Studii Economice din București (ASE); Academia Română, Institutul de Prognoză Economică (IPE))

Explorați vizual

- **Neuronpedia**
  - Platformă open-source pentru interpretabilitatea modelelor lingvistice: autoencoderele rare descompun activările unui strat în trăsături, iar pagina fiecărei trăsături arată textele pe care se activează cel mai puternic.
  - Comparați cele două perspective: SHAP și LIME atribuie o predicție intrărilor, pe cînd aceste trăsături descriu ce reprezintă interiorul modelului.

Lecturi și resurse suplimentare

- **C. Molnar, Interpretable Machine Learning**, gratuit online: christophm.github.io/interpretable-ml-book.
- **Biblioteca SHAP** și exemple: shap.readthedocs.io.
- **Rudin (2019)**: argumentele pentru modele interpretabile în deciziile cu miză mare.
- **Acest curs**: cod, date, notebook-uri — github.com/danpele/NEURAL_BIZ.
- **Încercați singuri**: notebooks/L11_notebook_ro.ipynb
  - Seminarul explică un model pentru German credit, cu o verificare a echității.
- **Replicarea studiului nostru de stabilitate**: github.com/QuantLet/GNN_XAI_Stability.

Bibliografie (1/3)

- Alvarez-Melis, D. & Jaakkola, T. S. (2018). On the robustness of interpretability methods. ICML Workshop on Human Interpretability in Machine Learning. arXiv:1806.08049
- Andrei, A.-V. & Pele, D. T. (2025). Explainable AI for electricity price anomaly detection: A SHAP-driven approach in Romania's energy market. *Preprints*. doi:10.20944/preprints202508.0410.v1
- Andrei, A.-V., Pele, D.-T. & Cramer, A.-A. (2026). Multi-horizon explainability in energy price forecasting: Disentangling market dynamics and renewable intermittency in Romania's liberalized electricity market. *Proceedings of the International Conference on Business Excellence*, 20(1), 104–117. doi:10.2478/picbe-2026-0014
- Andrei, A.-V., Pele, D.-T., Cramer, A.-A., Amza, A. & Diaconu, D. (2026). Explainable AI-based early warning system for electricity price spikes: Evidence from the Romanian market. *Proceedings of the International Conference on Business Excellence*, 20(1), 4820–4836. doi:10.2478/picbe-2026-0352
- Baesens, B., Roesch, D. & Scheule, H. (2016). *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*. Wiley.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. doi:10.1023/A:1010933404324
- Dal Pozzolo, A., Caelen, O., Johnson, R. A. & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. IEEE SSCI, 159–166. doi:10.1109/SSCI.2015.33
- Diaconu, D. & Pele, D. T. (2026). SHAP stability across random and forward-time validation in graph-augmented fraud detection. Submitted to *Expert Systems with Applications*. Code: github.com/QuantLet/GNN_XAI_Stability

Bibliografie (2/3)

- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. doi:10.1214/aos/1013203451
- Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65. doi:10.1080/10618600.2014.907095
- Kim, B. et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). ICML. arXiv:1711.11279
- Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. arXiv:1705.07874
- Lundberg, S. M. et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67. doi:10.1038/s42256-019-0138-9
- Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. KDD, 1135–1144. doi:10.1145/2939672.2939778
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. doi:10.1038/s42256-019-0048-x
- Samek, W., Binder, A., Montavon, G., Lapuschkin, S. & Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11), 2660–2673. doi:10.1109/TNNLS.2016.2599820

Bibliografie (3/3)

- Shapley, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games II*, 307–317. Princeton University Press.
- Simonyan, K., Vedaldi, A. & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. ICLR Workshop. arXiv:1312.6034
- Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. (2017). SmoothGrad: Removing noise by adding noise. arXiv:1706.03825
- Sundararajan, M., Taly, A. & Yan, Q. (2017). Axiomatic attribution for deep networks. ICML, 3319–3328. arXiv:1703.01365
- Wachter, S., Mittelstadt, B. & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887. doi:10.2139/ssrn.3063289
- Regulation (EU) 2024/1689 (Artificial Intelligence Act), *Official Journal of the European Union*, L, 12 July 2024. data.europa.eu/eli/reg/2024/1689/oj
- Board of Governors of the Federal Reserve System (2011). SR 11-7: Supervisory guidance on model risk management. federalreserve.gov

Anexa A: obținerea valorilor Shapley prin nucleul Shapley

- Se impune ca $g(\mathbf{z}) = \phi_0 + \sum_j \phi_j z_j$ să satisfacă $g(\mathbf{0}) = v(\emptyset)$ și $g(\mathbf{1}) = v(\{1..d\})$ (pondere infinită pe coaliția vidă și pe cea completă).
- Ecuațiile normale ale celor mai mici pătrate ponderate cu $\pi(\mathbf{z}) \propto (d-1)/[\binom{d}{|\mathbf{z}|}|\mathbf{z}|(d-|\mathbf{z}|)]$ au soluția unică $\phi_j =$ valoarea Shapley (Lundberg & Lee, 2017, teorema 2).
- Kernel SHAP eșantionează coaliții din π (împerecheind $\mathbf{z}$ cu $\mathbf{1}-\mathbf{z}$) și rezolvă regresia cu restricții
  - Eroarea scade ca $1/\sqrt{\text{eșantioane}}$.
- Pentru f liniar și variabile independente, $v(S) = \beta_0 + \sum_{j\in S}\beta_j x_j + \sum_{j\notin S}\beta_j \mathbb{E}x_j$, deci $\phi_j = \beta_j(x_j - \mathbb{E}x_j)$.

(Lundberg & Lee, 2017)

Anexa B: gradienții integrați ca metodă de integrare pe traiectorie

- Pentru un drum $\gamma(\alpha)$ de la $\mathbf{x}'$ la $\mathbf{x}$: $f(\mathbf{x}) - f(\mathbf{x}') = \int_0^1 \nabla f(\gamma(\alpha))^\top \gamma'(\alpha)\, d\alpha = \sum_j \int_0^1 \partial_j f(\gamma)\, \gamma_j'\, d\alpha$.
- Dreapta $\gamma(\alpha) = \mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}')$ dă $\gamma_j' = x_j - x_j'$: fiecare termen este IG_j, de unde completitudinea.
- Dintre metodele pe drum, dreapta este singura simetrică în raport cu variabilele (Sundararajan et al., 2017).
- Eroarea sumei Riemann cu punctul de mijloc este $O(m^{-2})$ pentru f neted
  - Punctele de frîngere ReLU o încetinesc, cum arată graficul de completitudine pentru m mic.

(Sundararajan, Taly & Yan, 2017)