Intel·ligència Artificial Explicable Aplicada a la Valoració de Crèdit

Joan Orteu Saiz

Treball Final de Grau, Grau en Matemàtiques

Director: Dr. Josep Vives

25 de febrer de 2024

Continguts

Continguts

Introducció

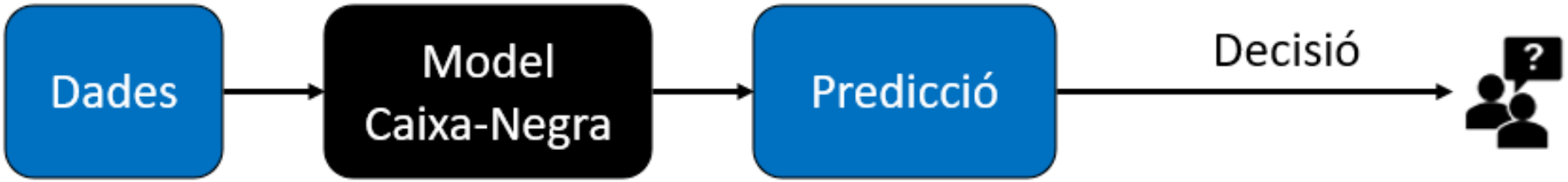


Figura: Estat actual de la Intel·ligència Artificial

Introducció

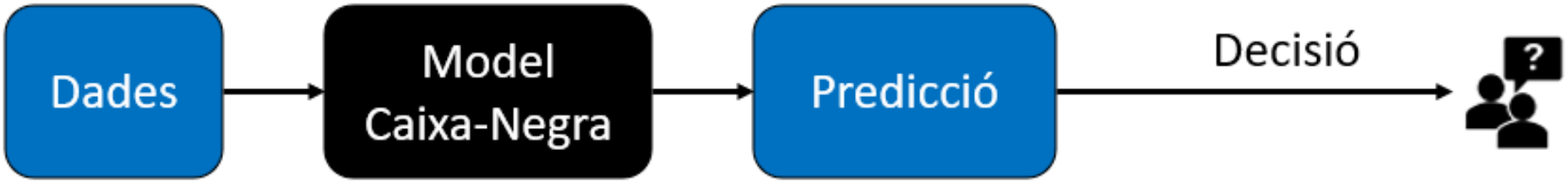


Figura: Estat actual de la Intel·ligència Artificial

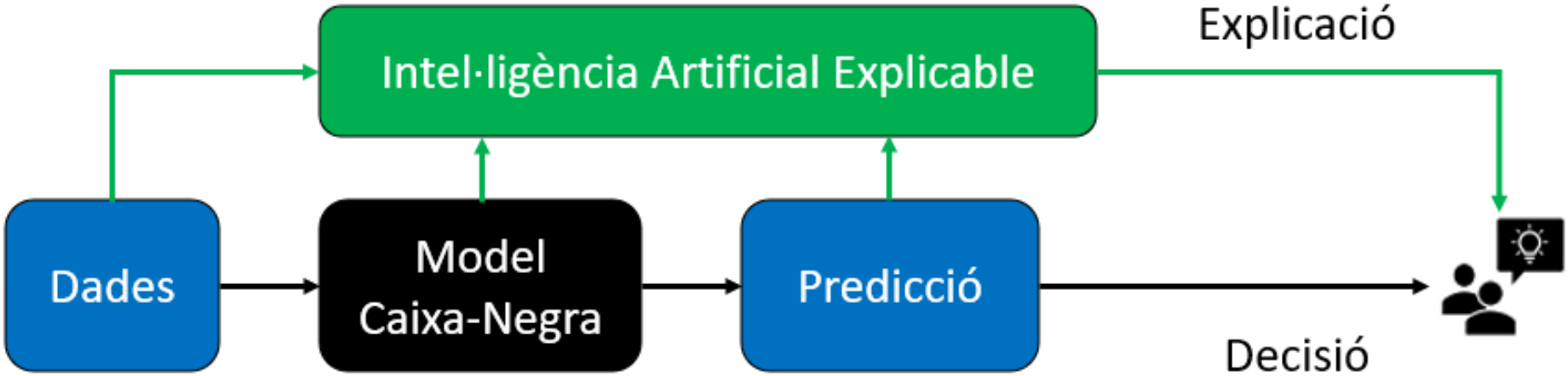


Figura: Intel·ligència Artificial (XAI)

Introducció

Introducció

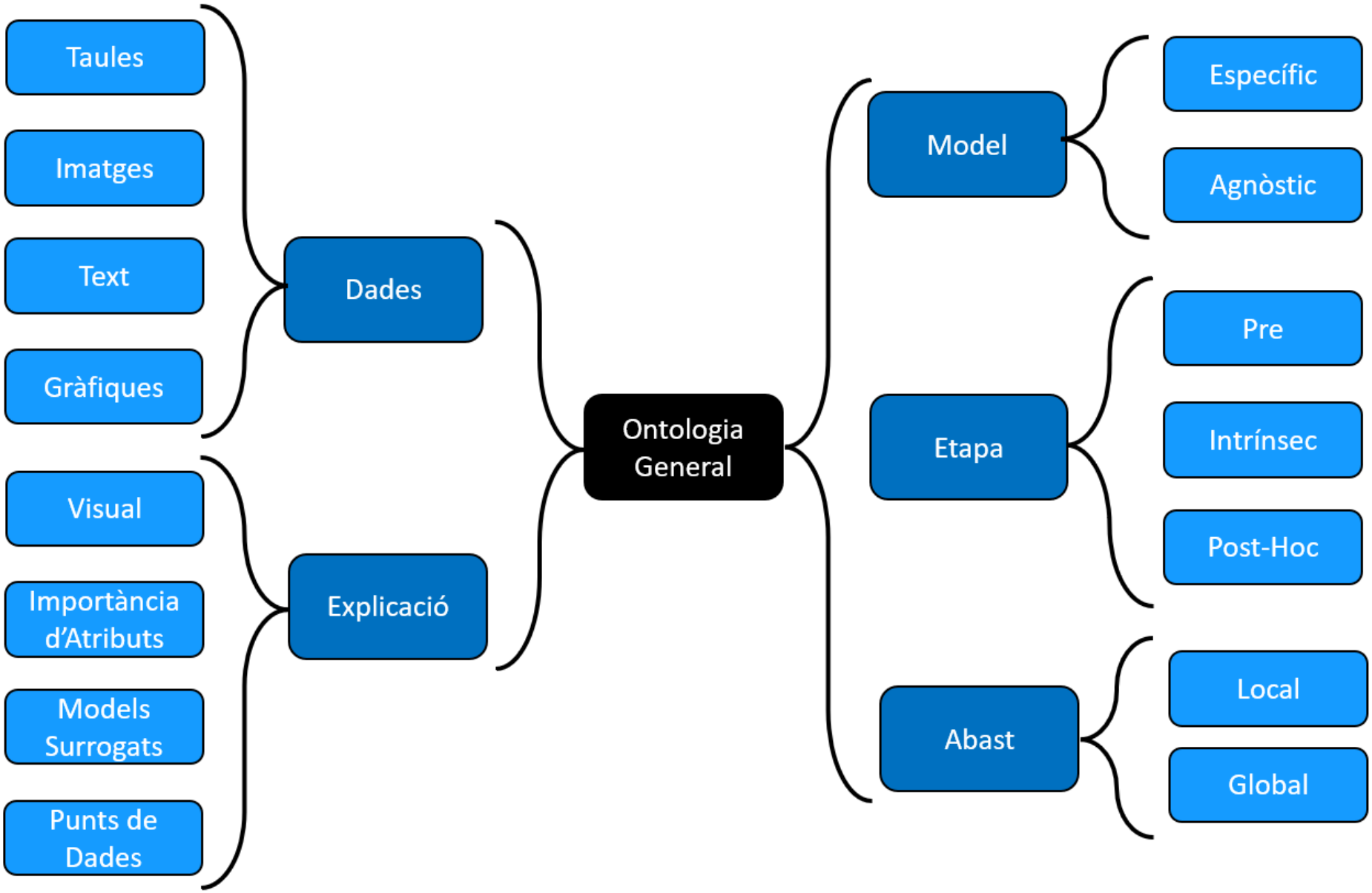
Continguts

Què és la Intel·ligència Artificial Explicable?

Definició de XAI (eXplainable Atificial Inteligence)

Conjunt de tècniques que podem utilitzar per a obtenir resposta a preguntes del tipus "per què" sobre models d'aprenentatge automàtic.

Taxonomia XAI



Taxonomia XAI

	SHAP	**LIME**
Etapa	Post-Hoc	Post-Hoc
Model	Agnòstic	Agnòstic
Abast	Local i Global	Local i Global
Dades	Totes	Totes
Explicació	Imp d'Atributs	Imp d'Atributs i Model Surrogat

SHAP

SHappley

Additive

ex**P**lanations

Història de SHAP

- **1953:** Lloyd Shapley, *"A Value for n-Persons Games"*.
- **2010:** Erik Štrumbelj i Igor Kononenko, *"An efficient explanation of individual classifications using game theory"*.
- **2016:** Marco Tulio Ribeiro i altres, *"Why should I trust you?"*.
- **2017:** Scott Lundberg i Su-In Lee, *"A unified Approach to Interpreting Model Predictions"*.

Jocs Cooperatius

S'estudia com es poden repartir els beneficis de una cooperació:

- Interès comú
- Interacció necessària entre jugadors
- Acord obligatori
- Benefici mutu

Definició de Jugadors

El **conjunt total de jugadors**, denotat per $N := \{1, ..., n\}$, és un conjunt on els seus elements són els agents prenedors de decisions d'un joc en el que tots estan inclosos. Utilitzarem indistintament conjunt total de jugadors i gran coalició.

Definició de Coalició

Per a cada subconjnt $S \subset N$, ens referim a S com a **coalició**.

Jocs d'Utilitat Transferible (Jocs-TU)

Definició de Jocs-TU

Un **joc d'utilitat transferible** és un parell (N, v), on N és el conjunt de jugadors i la funció característica del joc és $v : 2^N \to \mathbb{R}$ (on 2^N és el conjunt de parts de N). Per convenció, notem que la imatge del buit és zero ($v(\emptyset) := 0$).

Se li diu G^N a la classe dels jocs cooperatius amb n jugadors.

Sigui $(N, v) \in G^N$.

- Un jugador $i \in N$ es diu **jugador nul** si, per a qualsevol $S \subset N$, tenim que $v(S \cup \{i\}) - v(S) = 0$.
- **Dos jugadors** i, j **són simètrics** si, per a cada coalició $S \subset N \setminus \{i, j\}$, tenim que $v(S \cup \{i\}) = v(S \cup \{j\})$

Valor de Shapley

Definició de Valor de Joc

Se li diu **valor del joc** $v \in G^N$ a una funció ϕ de $\mathbb{R}^n$ en que $\phi(v)$ és un vector que representa en cada coordenada $\phi_i(v)$ el pagament o assignació que percep el jugador $i \in N$. En altres paraules, cada $\phi_i(v)$ és el Valor de Shapley que correspon al jugador i del joc v.

Definició de Valor de Shapley

El Valor de Shapley, Φ, es defineix per a cada $v \in G^N$ i cada $i \in N$ com:

$$\Phi_i(v) := \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!}(v(S \cup \{i\}) - v(S)). \qquad (1)$$

Valor de Shapley

Definició de Repartiment "Just"

Els axiomes proposats per Lloyd Shapley per definir un repartiment "just":

- **Eficiència:** $\forall v \in G^N$, tenim que $\sum_{i \in N} \phi_i(v) = v(N)$.
- **Jugador nul:** $\forall v \in G^N$ i $\forall i \in N$ jugador nul, tenim que $\phi_i(v) = 0$.
- **Simetria:** $\forall v \in G^N$ i $\forall i, j \in N$ de jugadors simètrics, tenim que $\phi_i(v) = \phi_j(v)$.
- **Additivitat:** $\forall v, w \in G^N$, tenim que $\phi(v + w) = \phi(v) + \phi(w)$.

Teorema de Shapley

El valor de Shapley és l'únic valor de joc en G^N que satisfà les propietats d'eficiència, jugador nul, simetria i additivitat simutàniament.

De Shapley a SHAP

Terme	**Concepte en ML**	**Terme matemàtic**
Jugador	Índex d'atribut	j
Número de Jugadors	Número d'atributs	N
Coalició	Conjunt d'atributs	$S \subseteq \{1, ..., N\}$
No en la Coalició	Atributs no en S	$C : C = \{1, ..., N\} \backslash S$
Mida de la Coalició	Número d'atributs	$\|S\|$
Utilitat de la gran coalició	Predicció per $x^{(i)}$ menys l'esperança de la predicció	$f(x^{(i)}) - \mathbb{E}(f(X))$
Utilitat de la coalició S	Predicció de la coalició S menys l'esperança del joc	$v_{f,x^{(i)}}(S)$
Valor de Shapley	Contribució de l'atribut j al pagament	$\phi_j^{(i)}$

De Shapley a SHAP

Definició de Valors SHAP

Donat un model f i instància $x^{(i)}$, la **funció de Valors SHAP** és la següent:

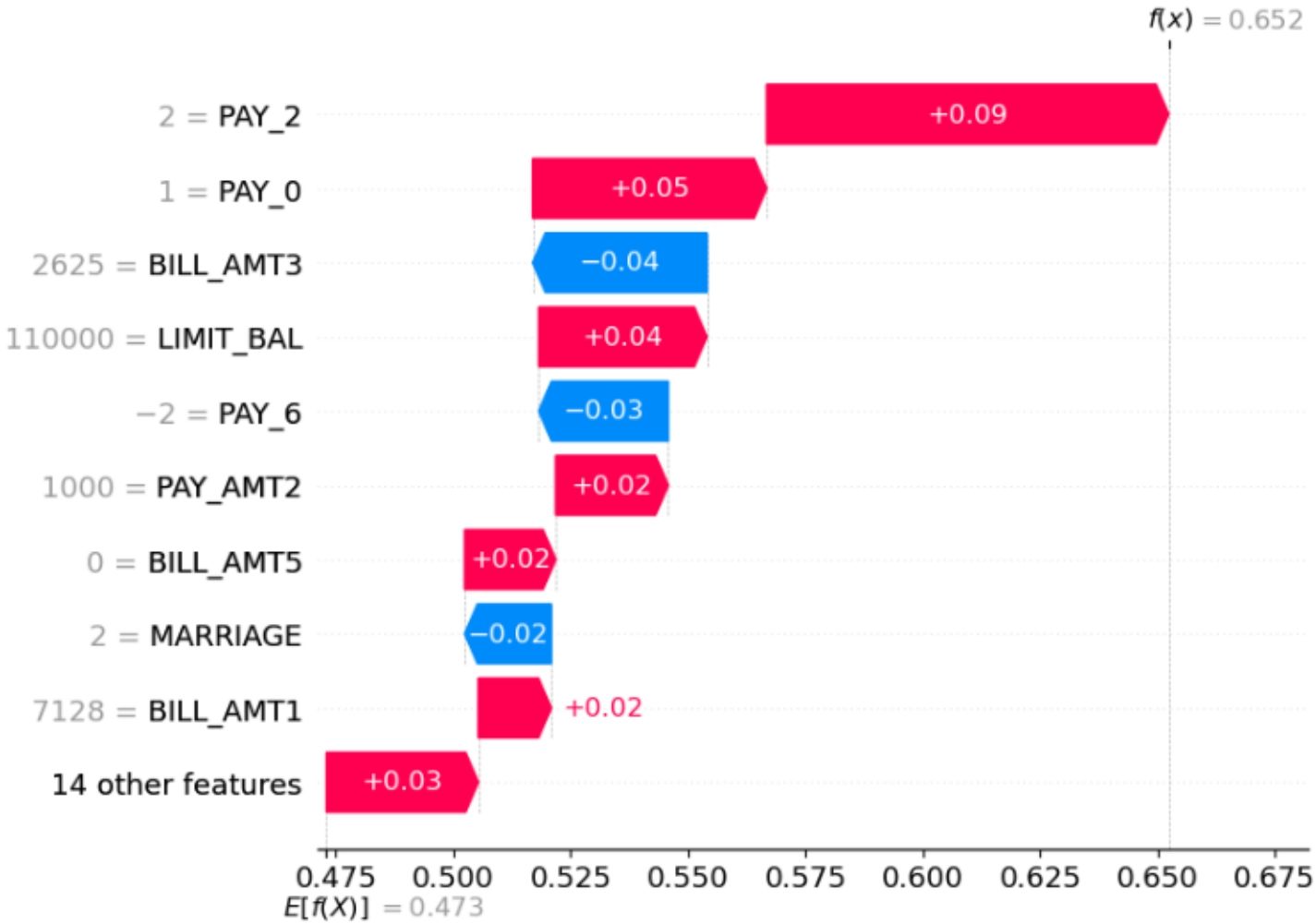
$$v_{f,x^{(i)}}(S) = \int f(x_S^{(i)} \cup X_C)\, d\mathbb{P}_{X_C} - \mathbb{E}(f(X)) \tag{2}$$

Definició de l'equació de SHAP

$$\phi_j^{(i)} = \sum_{S \subseteq \{1,\dots,N\} \setminus \{j\}} \frac{|S|!(N-|S|-1)!}{N!} \cdot \left(v_{f,x^{(i)}}(S \cup \{j\}) - v_{f,x^{(i)}}(S) \right). \tag{3}$$

on el **valor SHAP** $\phi_j^{(i)}$ d'un atribut d'un valor, és la contribució marginal mitjana de l'atribut j del valor $x^{(i)}$ a totes les possibles coalicions.

SHAP


f(x) = 0.652
2 = PAY_2
1 = PAY_0
2625 = BILL_AMT3
110000 = LIMIT_BAL
−2 = PAY_6
1000 = PAY_AMT2
0 = BILL_AMT5
2 = MARRIAGE
7128 = BILL_AMT1
14 other features
+0.09
+0.05
−0.04
+0.04
−0.03
+0.02
+0.02
−0.02
+0.02
+0.03
0.475
0.500
0.525
0.550
0.575
0.600
0.625
0.650
0.675
E[f(X)] = 0.473

Aproximacions SHAP

Integració de Montecarlo

La integració de Montecarlo, permet solucionar el problema que suposa no conèixer les distribucions. Utilitzant aquesta tècnica, l'aproximació a valors de SHAP és la següent:

$$\hat{v}(S) = \frac{1}{n} \sum_{k=1}^{n} (f(x_S^{(i)} \cup x_C^{(k)}) - f(x^{(k)})), \tag{4}$$

Mostreig de la Coalició

Ja que el nombre de coalicions augmenta exponencialment amb el nombre de característiques (2^N) i en els models d'aprenentatge autònom normalment s'utilitxen molts atributs, el temps de càlcul de totes les coalicions és prohibitiu. Una solució és fer Mostreig de les Coalicions, i per fer-ho es pot utilitzar la tècnica de permutacions.

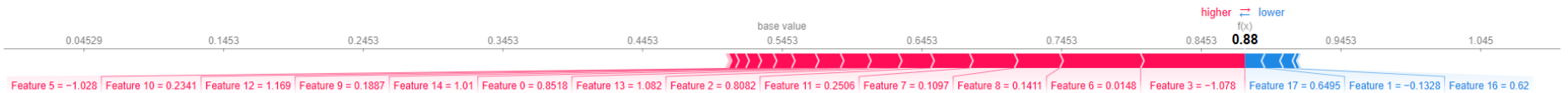
Avantatges i Limitacions

Avantatges

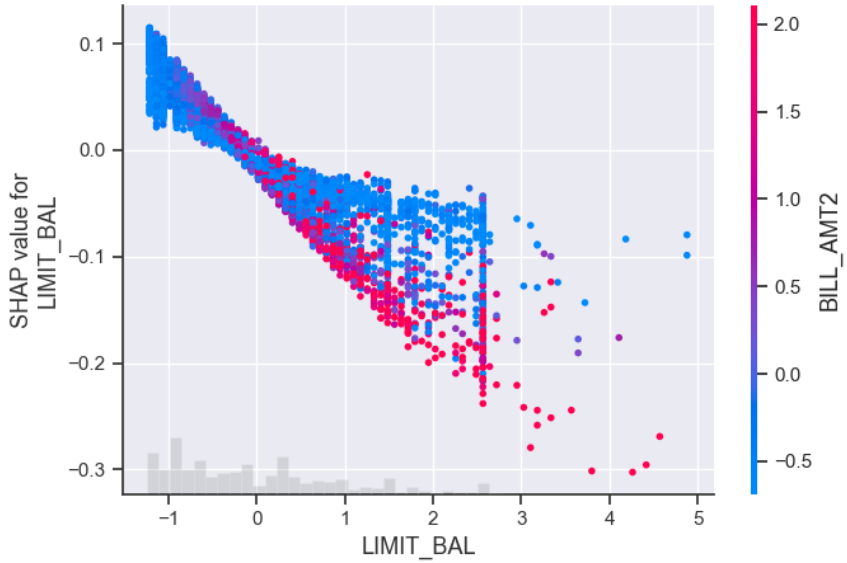
- Agnosticitat
- Taules, text i imatges
- Teoria sòlida
- Explicacions contrastables

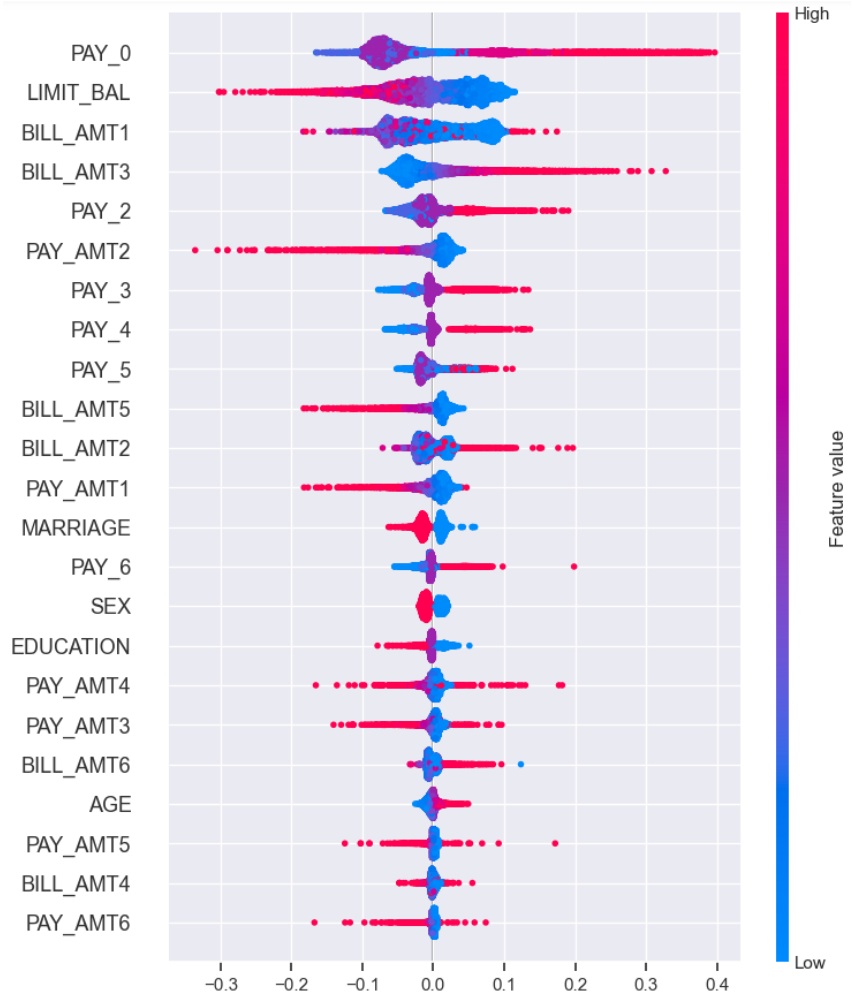
Limitacions

- Temps de càlcul excessiu
- Problema de la correlació
- Interaccions poden ser poc interpretables
- Valors de SHAP no habiliten acció
- Necessita accés a totes les dades



Visualitzacions SHAP





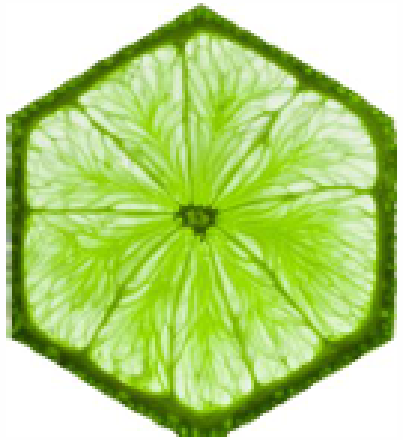
LIME

Local

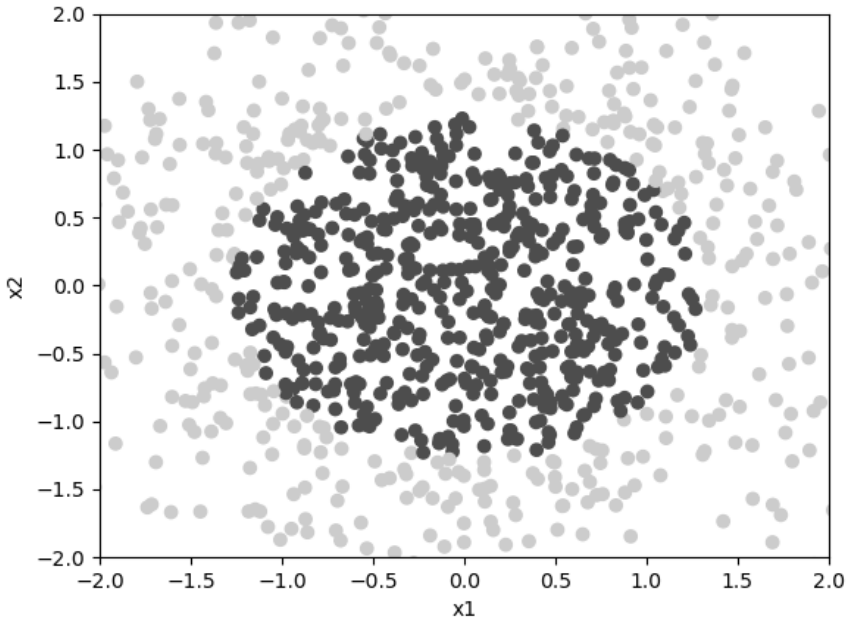
Interpretable

Model-agnostic

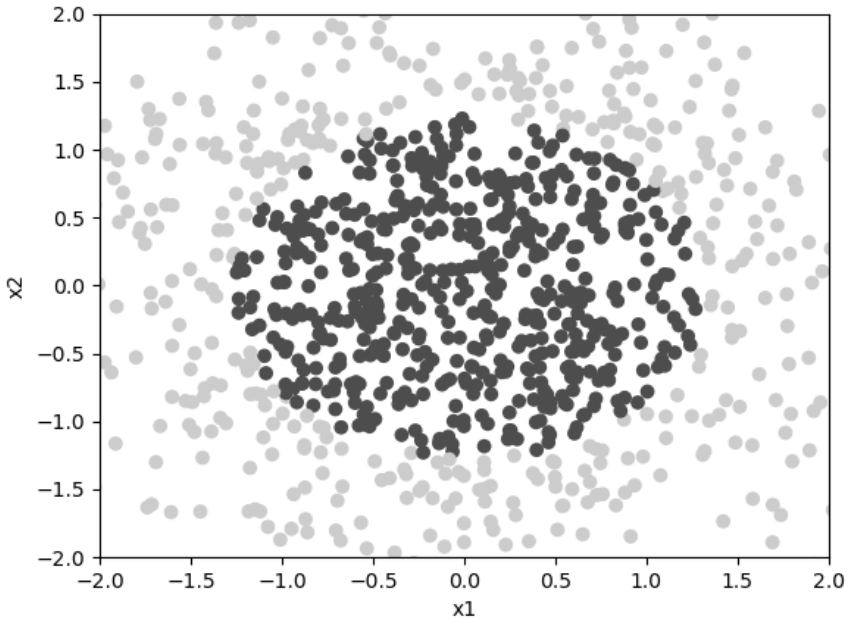
Explanations

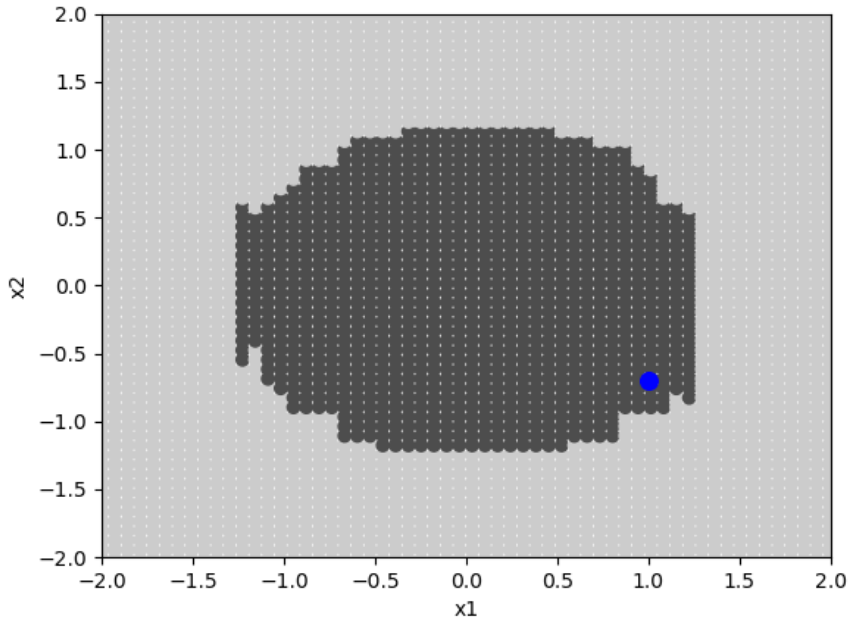


Exemple: Passos previs a aplicar LIME

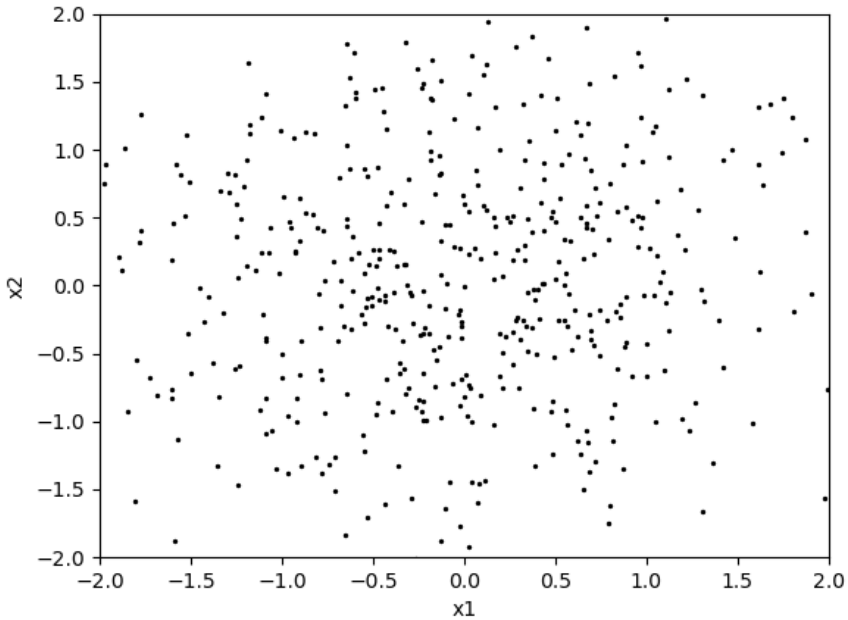


Exemple: Passos previs a aplicar LIME

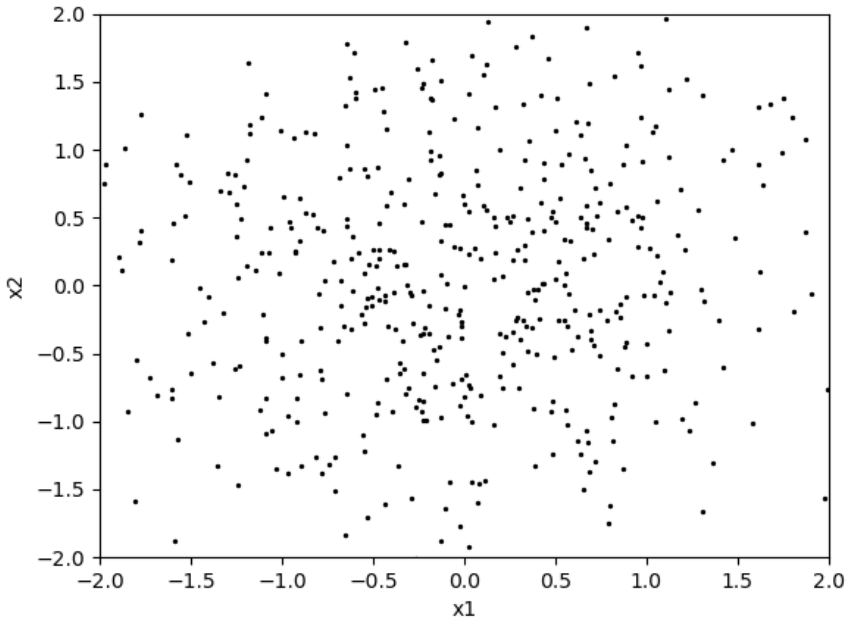


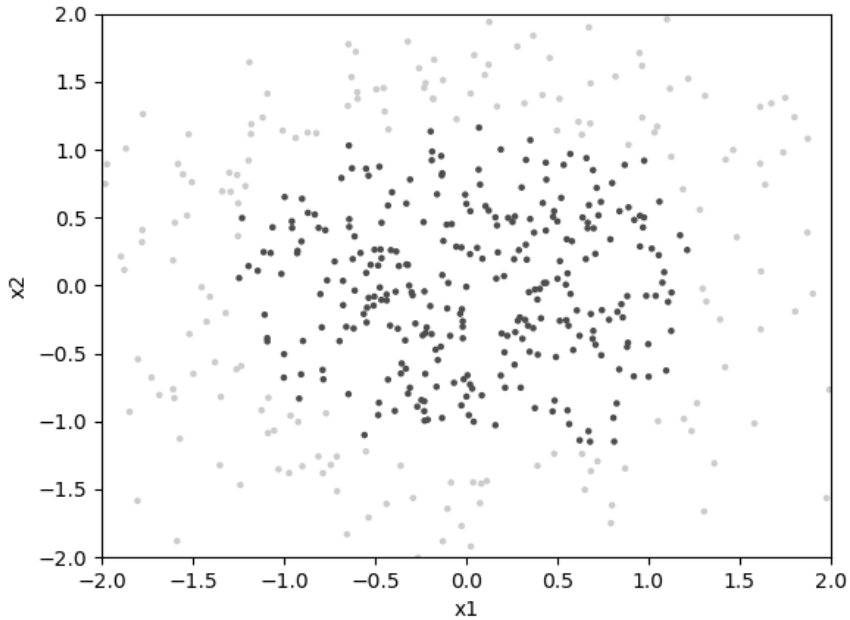


Exemple: Pas 1 i 2



Exemple: Pas 1 i 2





Exemple: Pas 3 i 4

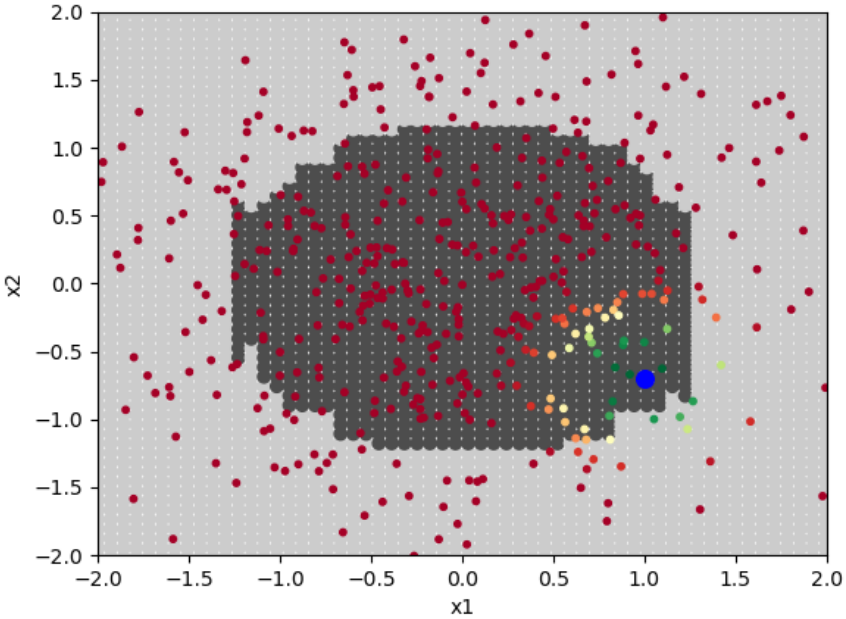


Figura: Pertorbacions aleatòries al voltant de la instància a explicar evaluades i pesades.

Exemple: Pas 3 i 4

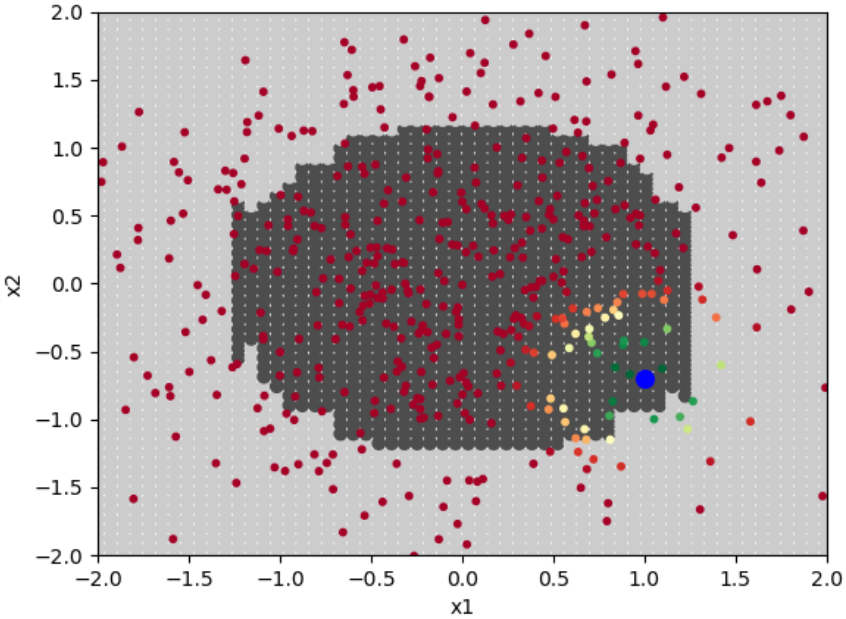


Figura: Pertorbacions aleatòries al voltant de la instància a explicar evaluades i pesades.

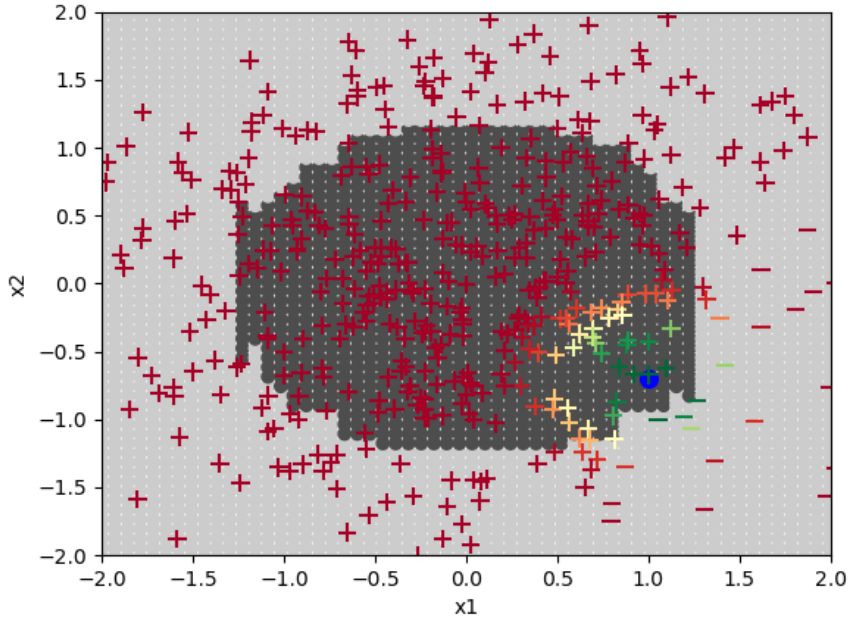


Figura: Pertorbacions aleatòries al voltant de la instància a explicar evaluades, pesades i amb l'explicació final.

Problema d'optimització LIME

Explicació LIME

Sigui G la classe de models d'aprenentatge automàtic intrínsecament interpretables, h la Caixa Negra que es vol explicar, x la instància a explicar, i π_x la mesura de proximitat de x.
L'explicació LIME $\xi(x)$ en el punt x es troba minimitzant $\mathcal{L}$, mantenint prou petit $\Omega(g)$:

$$\xi(x) = \arg\min_{g \in G} \left(\mathcal{L}(h, g, \pi_x) + \Omega(g)\right) \tag{5}$$

- $\mathcal{L}(h, g, \pi_x)$ mesura la precisió amb la que g s'aproxima a h, utilitzant la mesura π_x.
- $\Omega(g)$ mesura la complexitat del model interpretable g.

Problema d'optimització LIME

Mostreig per l'Exploració Local de LIME

No es vol fer cap suposició sobre h, ja que es vol que sigui independent del model. Per tant, es generen instàncies pertorbades aleatòries $z_1, \ldots, z_q \in \mathbb{R}^d$. Per dades tabulars, LIME crea noves instàncies pertorbant individualment cada característica de x a partir duna distribució normal.

Terme de pèrdua de LIME

A LIME, es selecciona la família de models intèrprets G com la classe de models lineals $g(z) = w_g \cdot z$. Aquest tipus de model es fa servir per minimitzar una regressió lineal ponderada:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z \in Z} \pi_x(z)(h(z) - g(z))^2 \tag{6}$$

On la funció del pes π_x, és un *nucli exponencial* definit sobre alguna mesura de distància $D(x, z)$: $\pi_x(z) = \exp\left(-D(x, z)^2/\sigma^2\right)$

Algorisme SP-LIME

Equació de cobertura

Sigui B el nombre màxim d'instàncies que un individu pot interpretar. Sigui W la matriu de les importàncies d'atributs de les instàncies preseleccionades. Sigui I_j la importància global de cada atribut.
Es vol trobar el subgrup V de X que tingui cobertura màxima de la base de dades V tq $|V| \geq B$.

$$\text{Pick}(W, I) = \underset{V, |V| \leq B}{\text{argmax}}\, c(V, W, I) \tag{7}$$

On $c(V, W, I)$ és la cobertura d'un subgrup V:

$$c(V, W, I) = \sum_{j=1}^{d'} \mathbf{1}_{[\exists i \in V : W_{ij} > 0]} I_j \tag{8}$$

Avantatges i Limitacions

Avantatges

- Agnosticitat
- Taules, text i imatges
- Concís i comprensible
- Model surrogat

Limitacions

- Definició de veïnat
- No linearitat
- Instàncies improbables
- Explicacions inestables

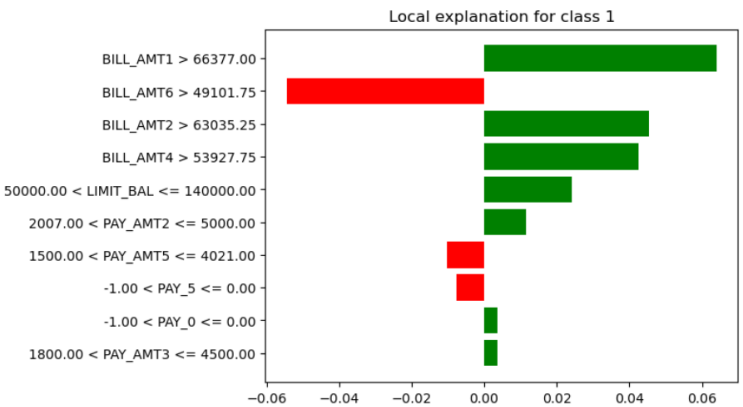


Figura: Explicació LIME d'una instància.

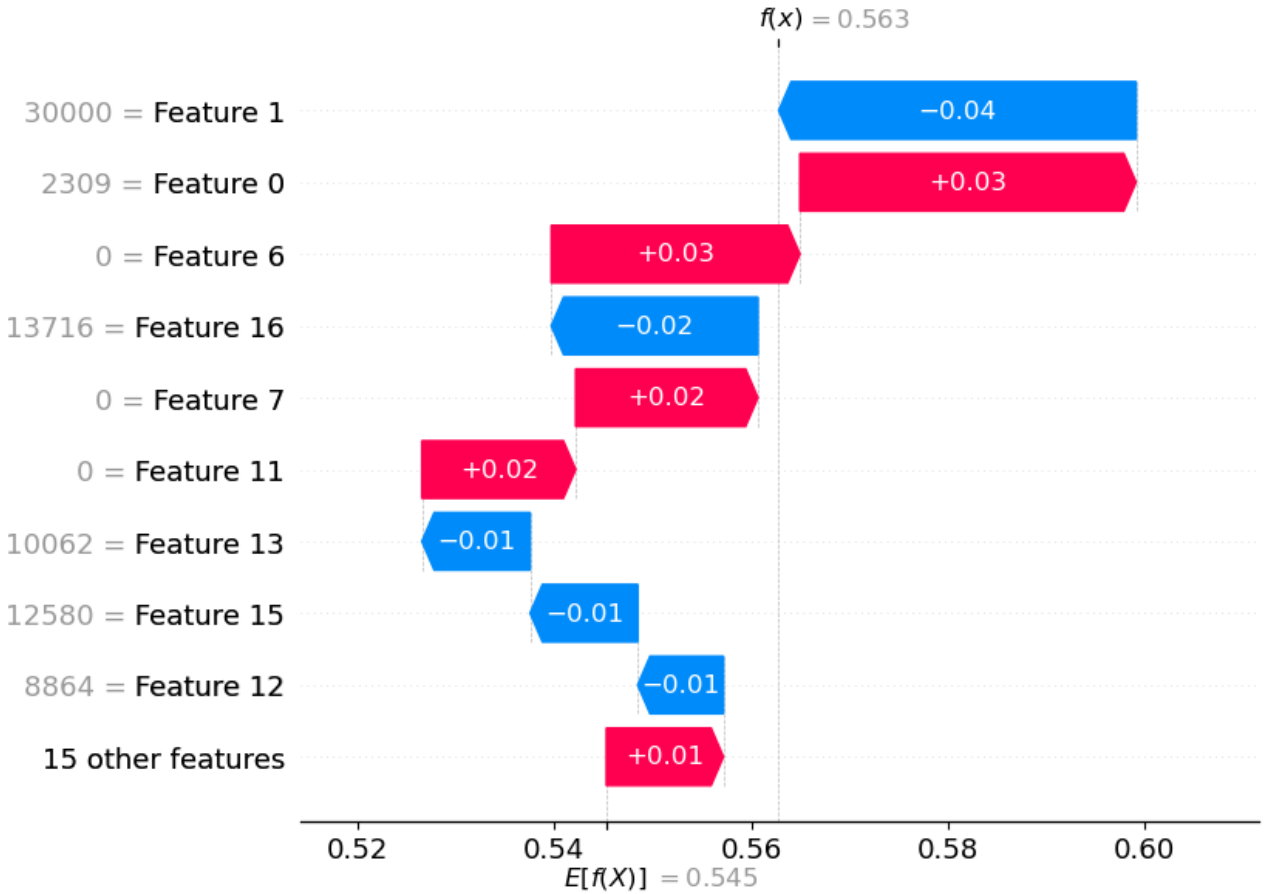
	Rule	Value	Importance
0	BILL_AMT1 > 66377.00	70122	0.063998
1	BILL_AMT6 > 49101.75	70212	-0.054412
2	BILL_AMT2 > 63035.25	69080	0.045362
3	BILL_AMT4 > 53927.75	69753	0.042438
4	50000.00 < LIMIT_BAL <= 140000.00	70000	0.024244
5	2007.00 < PAY_AMT2 <= 5000.00	3112	0.011432
6	1500.00 < PAY_AMT5 <= 4021.00	2500	-0.010324
7	-1.00 < PAY_5 <= 0.00	0	-0.007640
8	-1.00 < PAY_0 <= 0.00	0	0.003741
9	1800.00 < PAY_AMT3 <= 4500.00	3000	0.003681

SHAP vs LIME

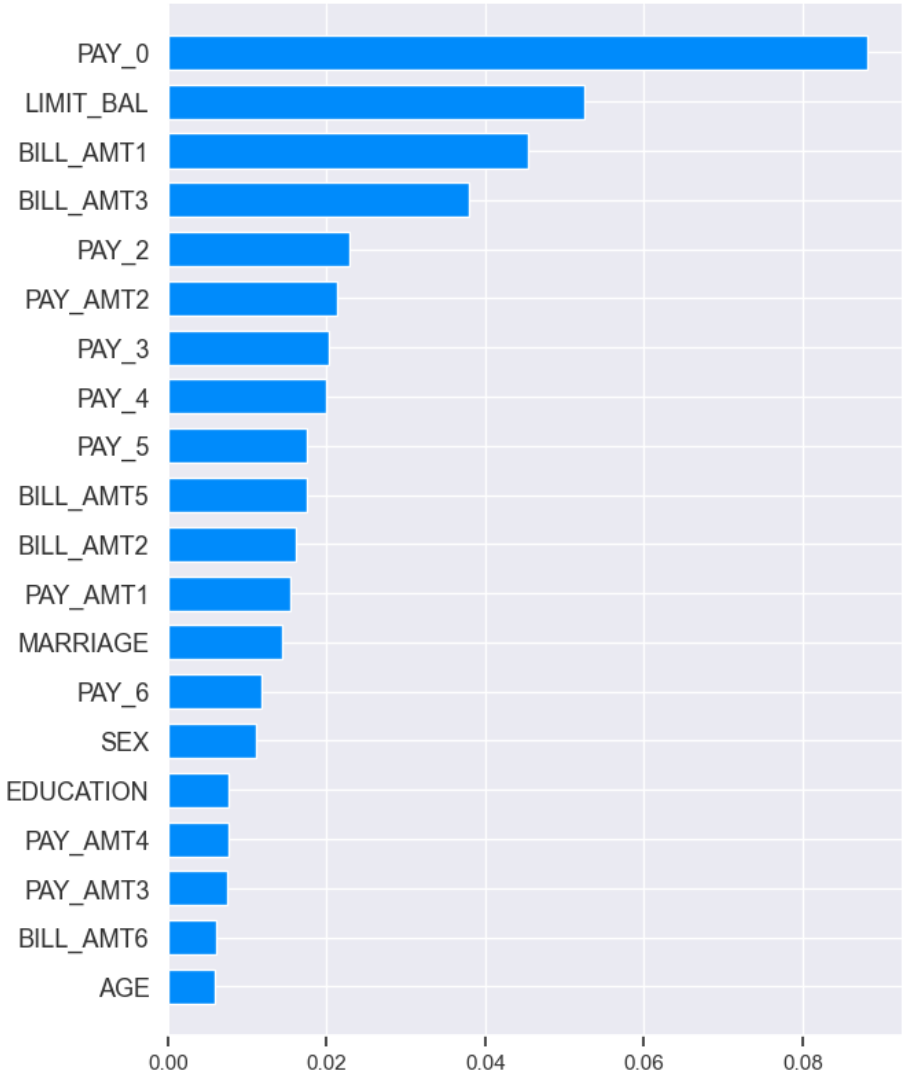
	SHAP	LIME
Distribució eficient	✓	✓
Garantia d'explicació	✓	✗
Velocitat de càlcul	✗	✓
Accés a dades	✗	✓
Complexitat d'explicacions	✗	✓
Instàncies improvables	✗	✗
Explicacions habiliten acció	✗	✓

Continguts

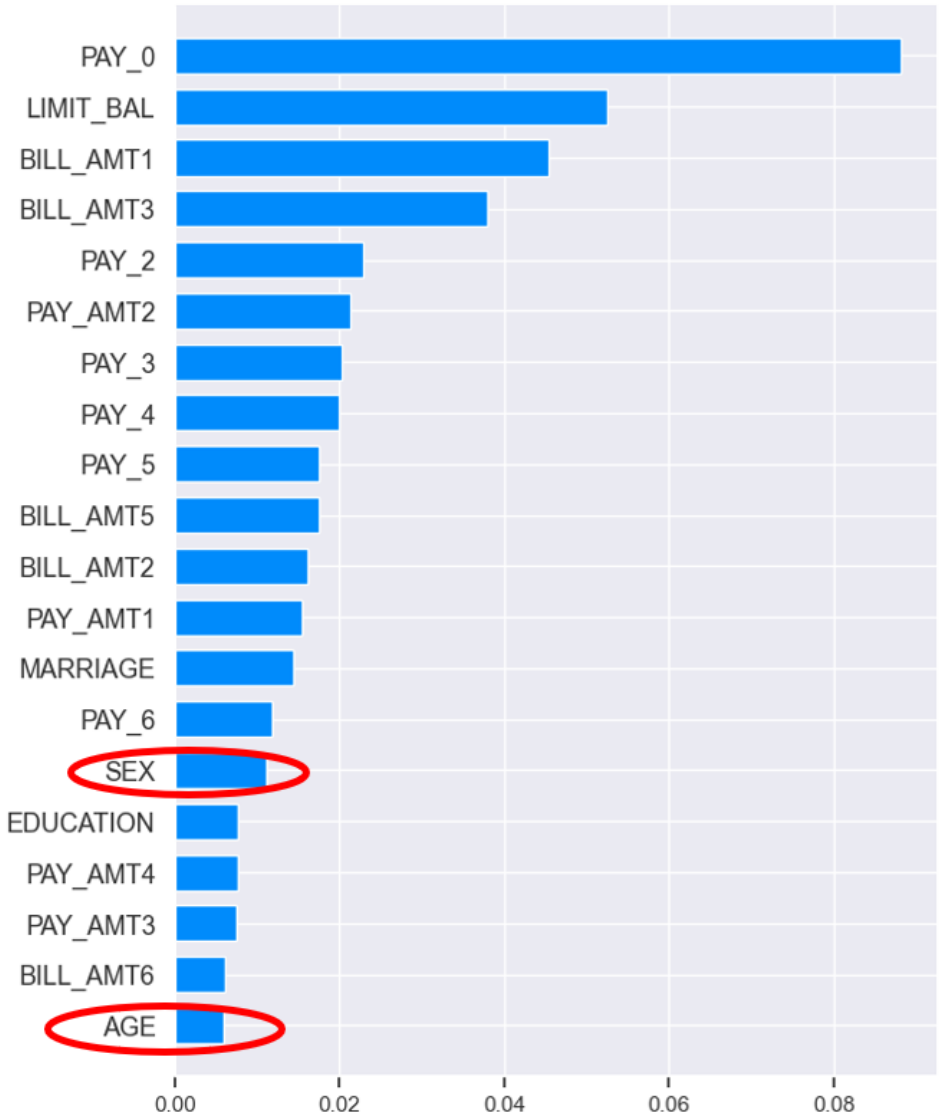
Experiment 1



Experiment 2



Experiment 2



Experiment 2

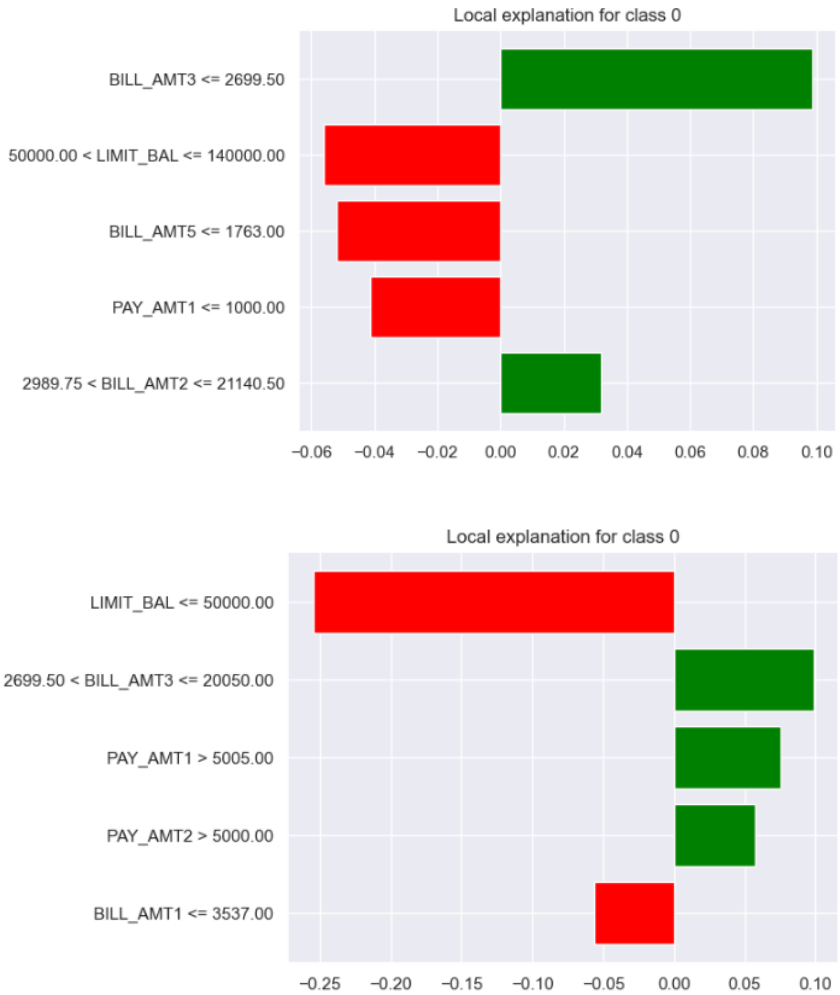
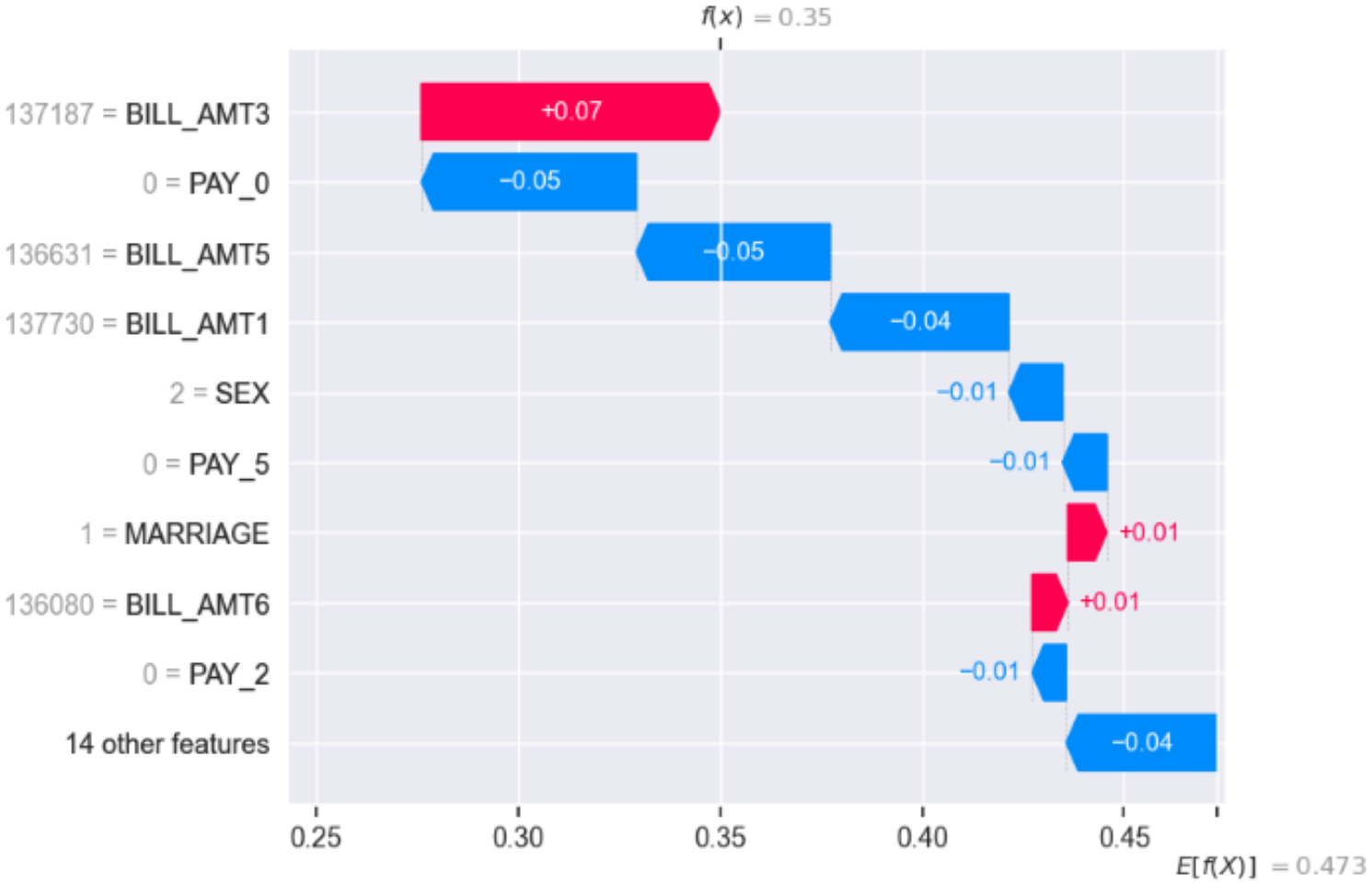


Figura: Visualització de l'algorisme SB-LIME sobre el model ANN.

Experiment 3



Continguts

Conclusió

- El camp de XAI té un gran potencial.
- SHAP i LIME són molt útils, però tenen defectes.

Treball Futur

- Estudi d'aproximacions de SHAP.
- Estudi d'interaccions entre atributs.
- Comparar resultats amb altres tècniques de XAI.