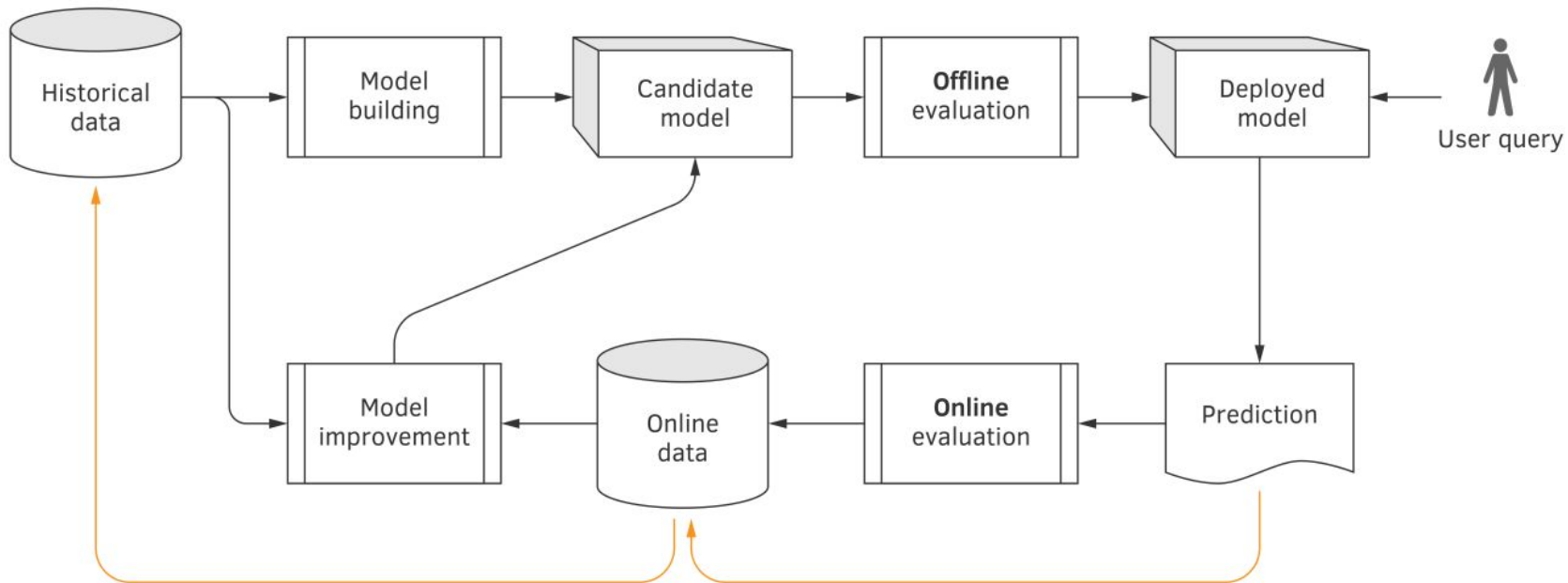


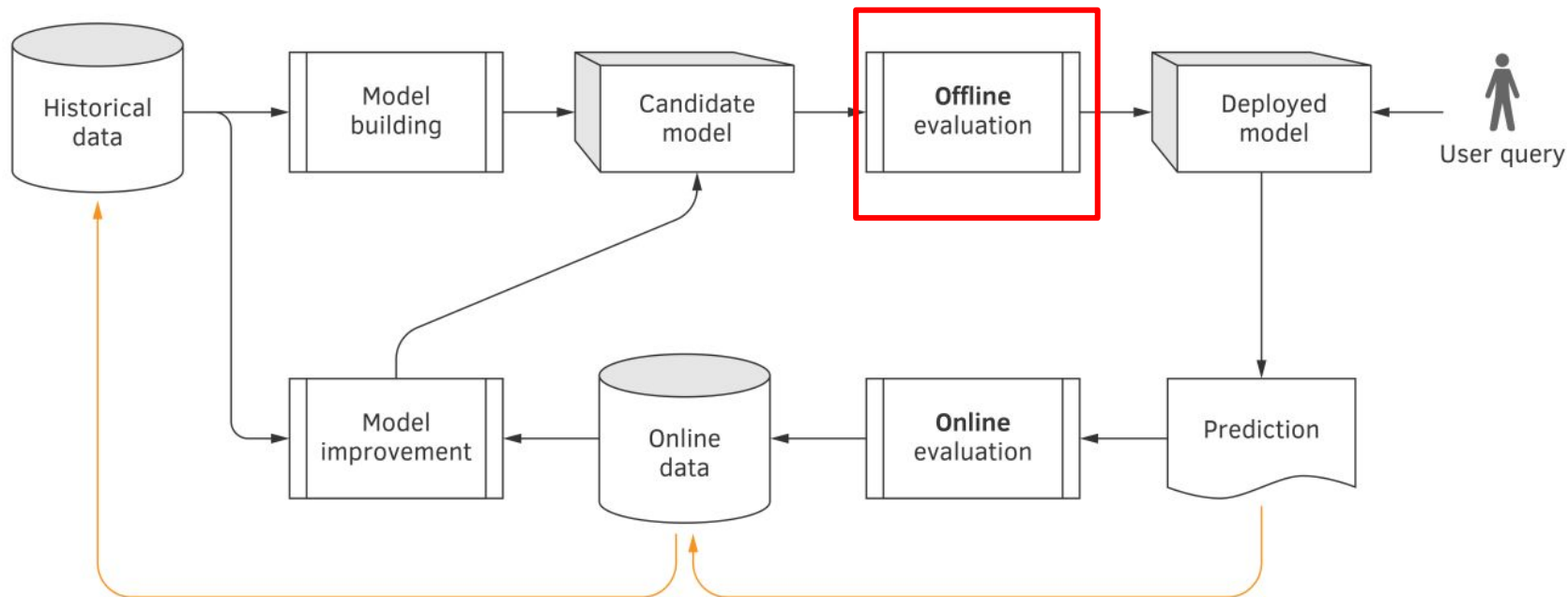
Fondamenti MLOps:

parte 3

Debugging dei modelli



Debugging dei modelli



Programma di oggi

Debugging e analisi dei modelli → Da pre-produzione a produzione

Tramite:

1. Utilizzo tecniche **eXplainable AI** (XAI)
2. Applicazione di concetti fondamentali di **analisi di incertezza** in ambito machine learning.

Possibili problemi legati ai modelli

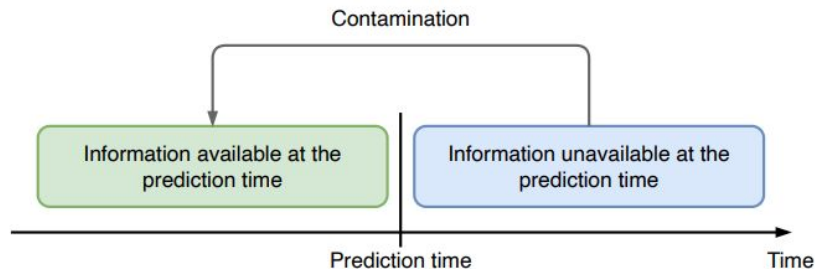
(che affronteremo oggi)

- 1) Data leakage
- 2) Robustezza modelli
- 3) Bias e fairness

Data leakage

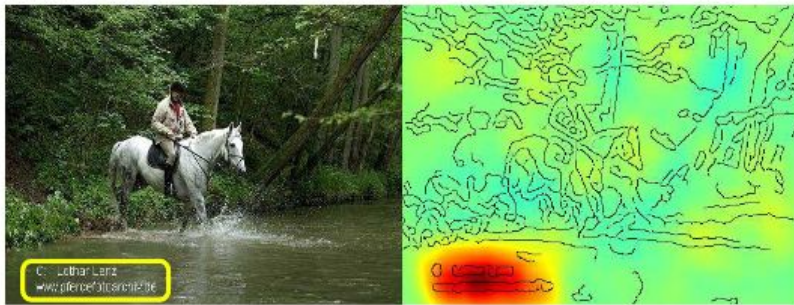
Il modello viene fornito in fase di allenamento di dati che non sono presenti in fase di inferenza. Motivi tipici:

- Una delle features 'nasconde' il target
- Una delle features viene dal 'futuro'



Data leakage

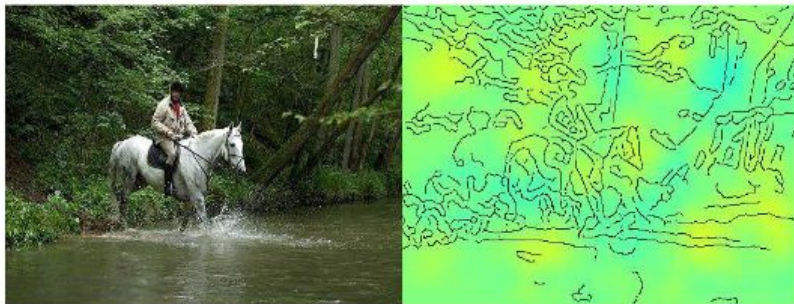
Horse-picture from Pascal VOC data set



Source tag
present



Classified
as horse



No source
tag present



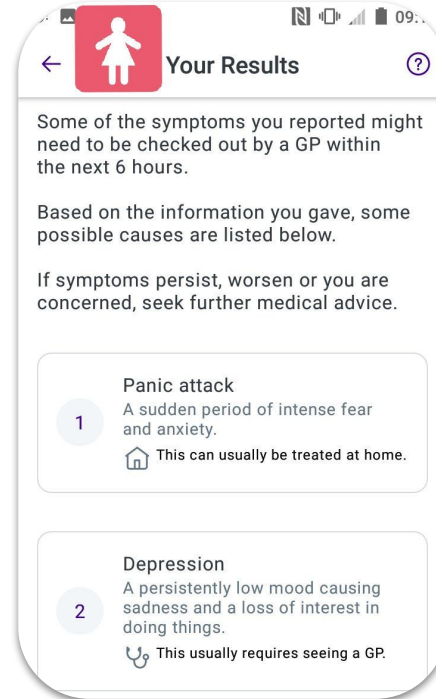
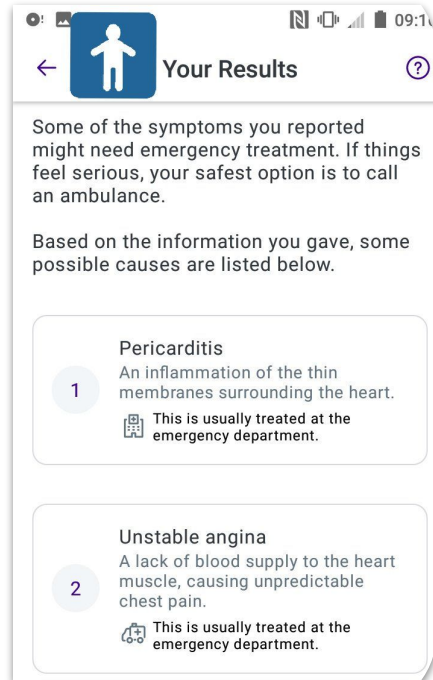
Not classified
as horse

Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. Lapuschkin et al. Nature comm (2019)

Bias e Fairness

Modelli imparano da dati → Dati possono contenere diversi tipi di bias che potrebbero non risultare accettabili in un modello in produzione.

Bias e Fairness

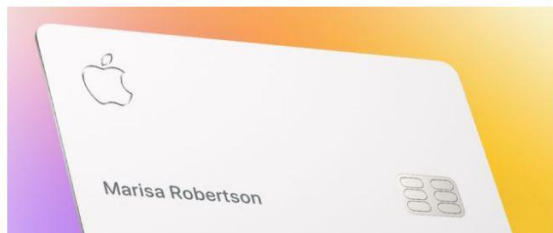


Bias e Fairness

Apple's 'sexist' credit card investigated by US regulator

11 November 2019

f t e Share



Steve Wozniak ✓
@stevewoz

Replying to @dhh

The same thing happened to us. We have no separate bank accounts or credit cards or assets of any kind. We both have the same high limits on our cards, including our AmEx Centurion card. But 10x on the Apple Card.

7:58 AM · Nov 10, 2019 · Twitter Web App



DHH ✓
@dhh

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

9:34 PM · Nov 7, 2019 · Twitter for iPhone

12.8K Retweets 28.6K Likes



DHH ✓ @dhh · Nov 7, 2019
Replying to @dhh

I'm surprised that they even let her apply for a card without the signed approval of her spouse? I mean, can you really trust women with a credit card these days??!

86 270 4.4K

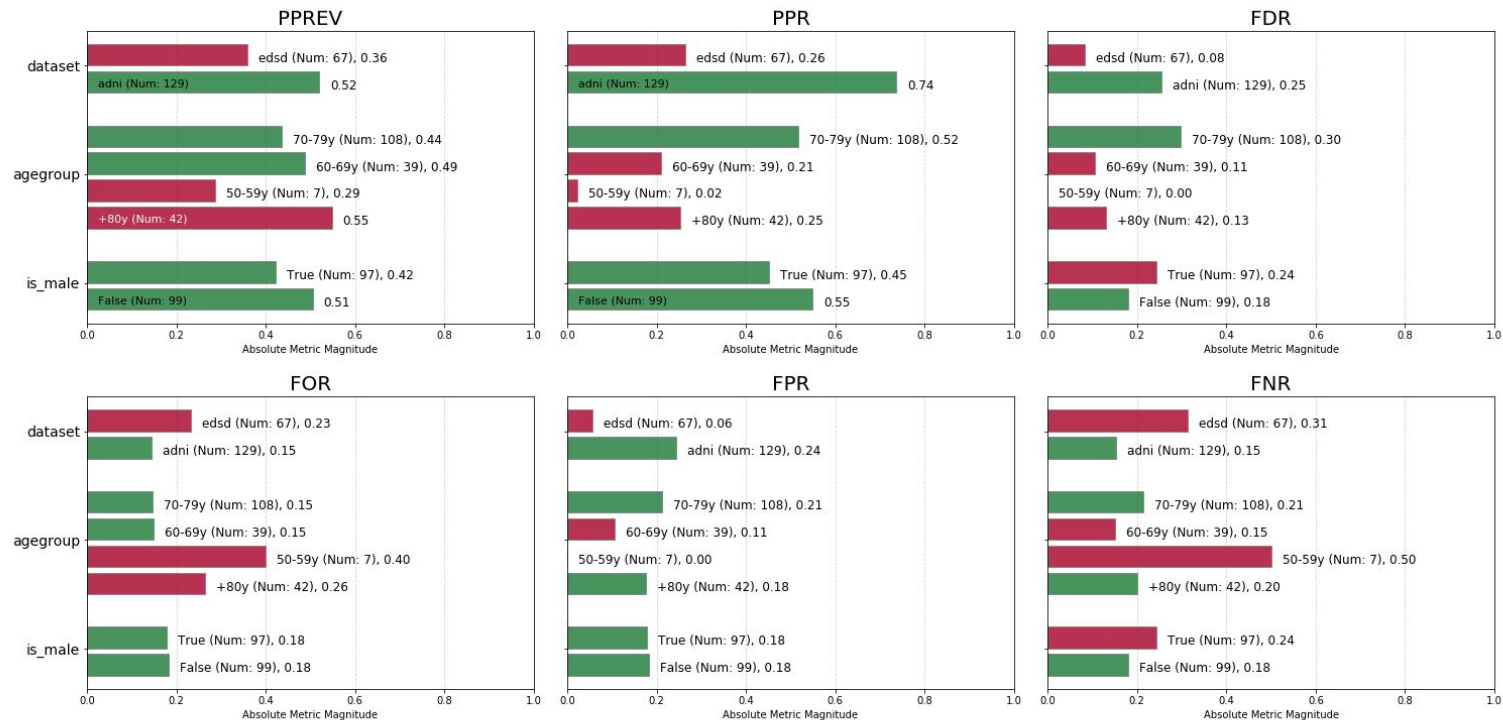


DHH ✓ @dhh · Nov 7, 2019

It gets even worse. Even when she pays off her ridiculously low limit in full, the card won't approve any spending until the next billing period. Women apparently aren't good credit risks even when they pay off the fucking balance in advance and in full.

Bias e Fairness

Aequitas toolkit



Robustezza modelli

Robustezza di un modello definita come la **stabilità rispetto a piccole perturbazioni**.

In determinati contesti modelli poco robusto possono avere conseguenze legate alla sicurezza.

Es: Modello in ambito computer vision che commette errori quando in presenza di artefatti legati alla compressione di un'immagine

Robustezza modelli

Robustezza di un modello definita come la **stabilità rispetto a piccole perturbazioni**.

In determinati contesti modelli poco robusto possono avere conseguenze legate alla sicurezza.

Es: Modello in ambito computer vision che commette errori quando in presenza di artefatti legati alla compressione di un'immagine

$$\|\mathbf{x} - \mathbf{x}'\| \leq \delta$$



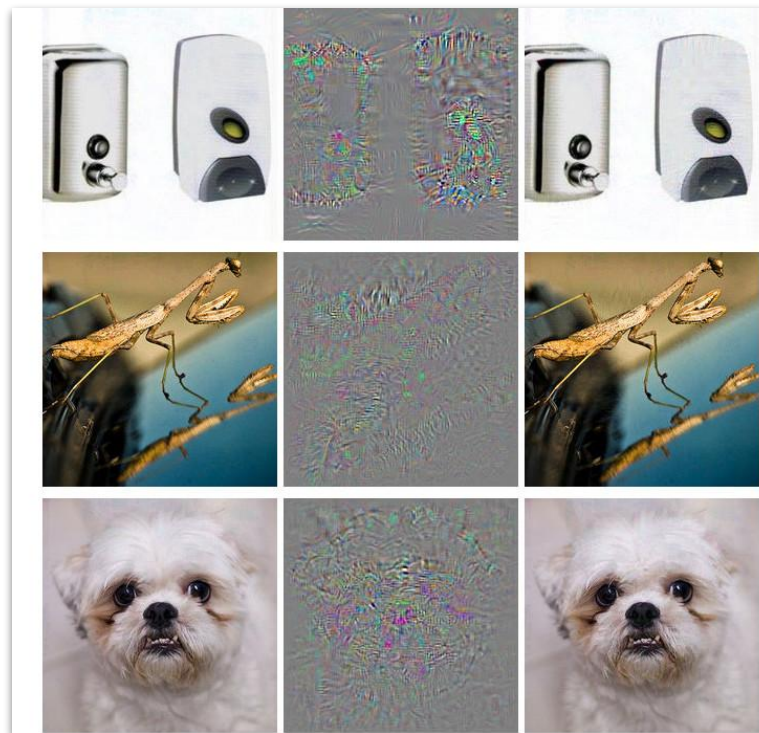
$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq \epsilon$$

Adversarial robustness

Fenomeno legato alla robustezza di un modello → Perturbazioni **mirate** che portano a errori importanti.



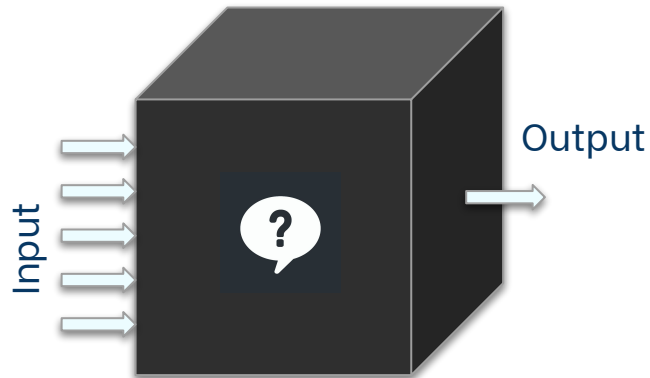
Source: Berkeley AI Research (BAIR)



“Intriguing properties of neural networks”, Szegedy et.al, 2013

Spiegabilità modelli


il problema della black-box



Fenomeni descritti fino ad ora aumentano con l'aumentare della **complessità** di un modello.

L'unica maniera per intercettarli in fase di debugging e' usando tecniche di **interpretazione** di modelli.

Spiegabilità modelli oltre al debugging

Engineering Topics ▾ Special Reports ▾ Blogs ▾ Multimedia ▾ The Magazine ▾ Professional Resources ▾ Search ▾

Feature | Biomedical | Diagnostics


02 Apr 2019 | 15:00 GMT

How IBM Watson Overpromised and Underdelivered on AI Health Care

After its triumph on Jeopardy!, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

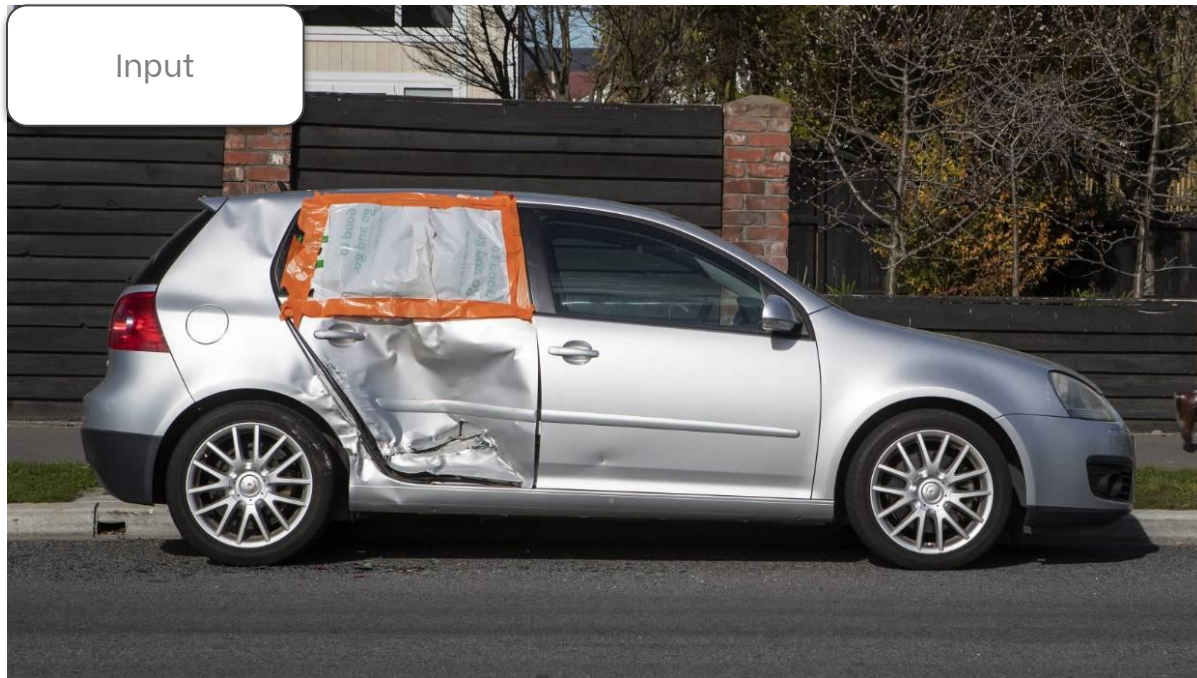
By **Eliza Strickland**

In 2014, IBM opened swanky new headquarters for its artificial intelligence division, known as **IBM Watson**. Inside the glassy tower in lower Manhattan, IBMers can bring prospective clients and visiting journalists into the “immersion room,” which resembles a miniature planetarium. There, in the darkened space, visitors sit on swiveling stools while fancy graphics flash around the curved screens covering the



Spiegabilità modelli oltre al debugging

Input



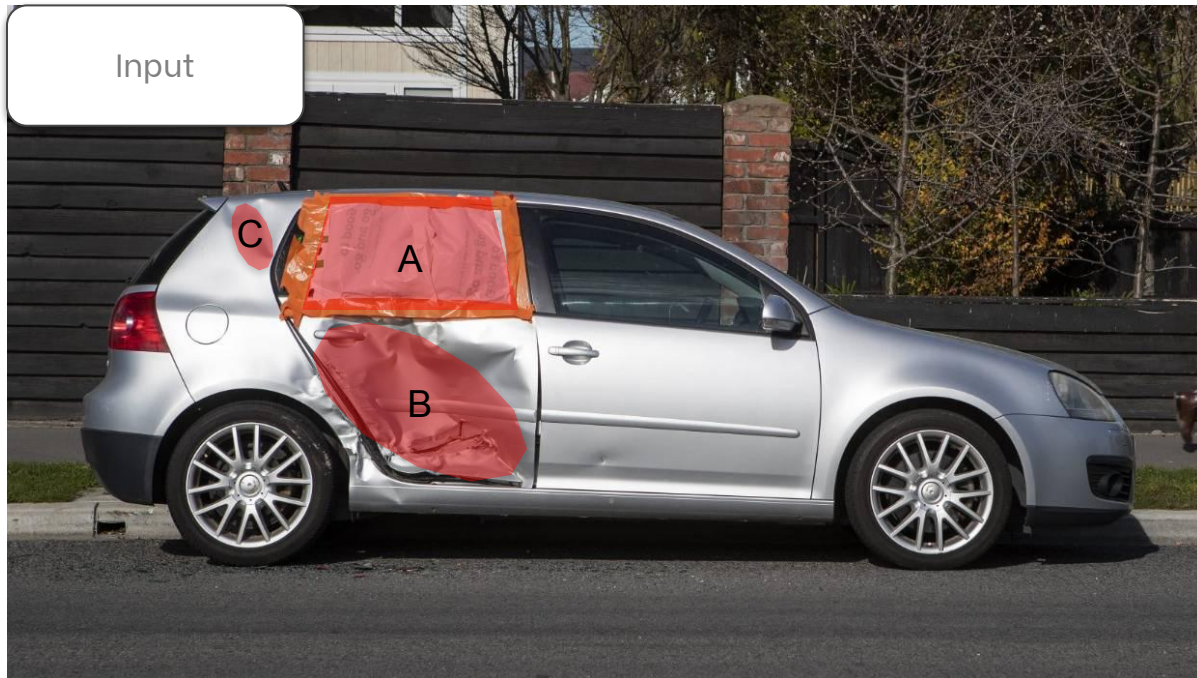
Output:

Damage= 2000€

Spiegabilità modelli

oltre al debugging

Input

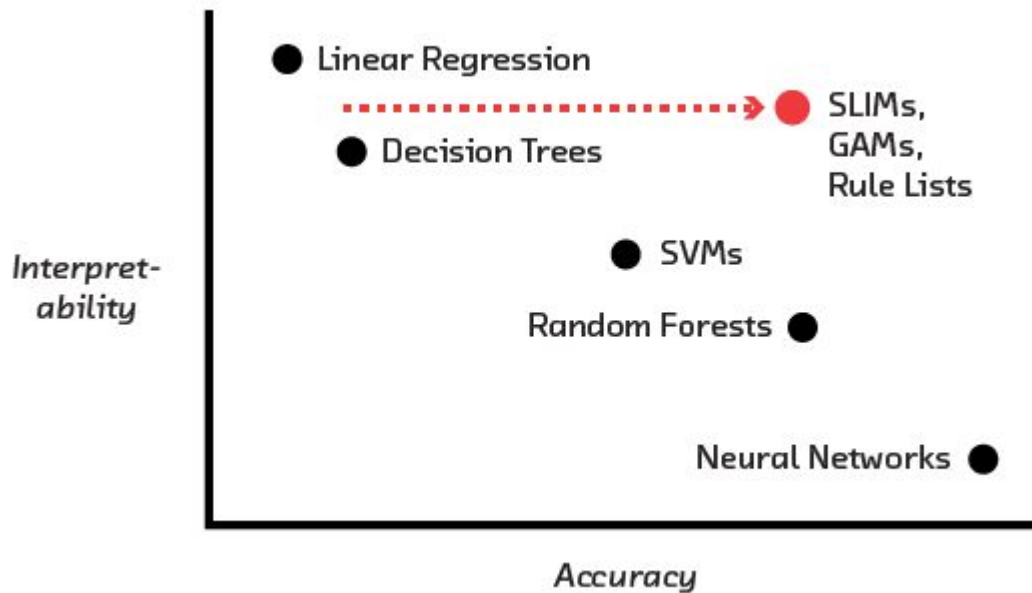


Output:
Damage = 2000€
Breakdown:
A = 300€
B = 1200€
C = 500€

Historical Example for B
Label = 2500€

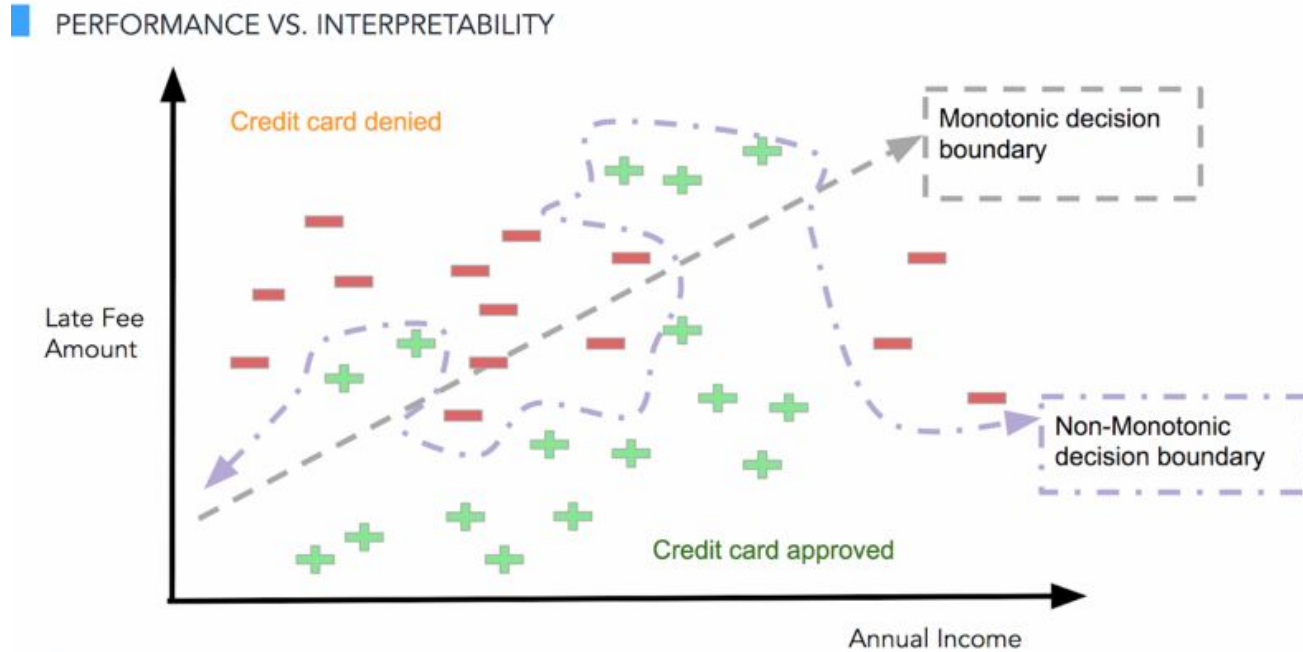


Spiegabilità modelli



<https://ff06-2020.fastforwardlabs.com/>

Tassonomia eXplainable AI

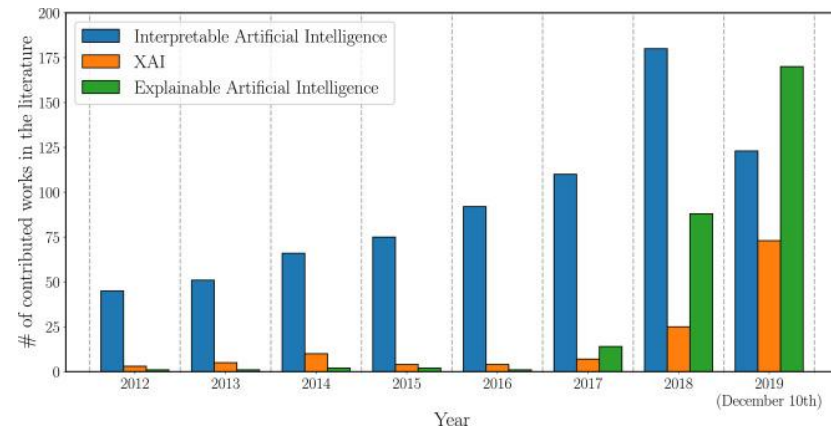


<https://www.kdnuggets.com/2018/12/explainable-ai-model-interpretation-strategies.html>

L'eXplainable AI (XAI)

Ambito di ricerca il cui scopo e' quello di aiutare umani a **interpretare** decisioni prese da modelli.

Interesse accademico e' in costante crescita → Sempre più metodi e librerie.

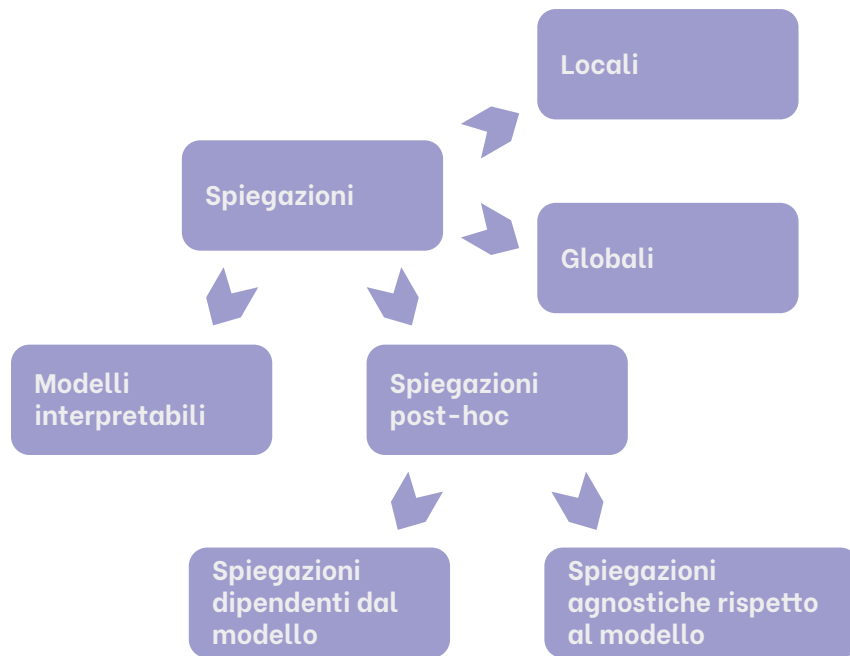


Arrieta et al, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, 58, 2020

Tassonomia eXplainable AI

Modi di spiegare i modelli possono essere molteplici.

Nelle prossime slides ci focalizzeremo su spiegazioni **locali**, agnostiche e non.

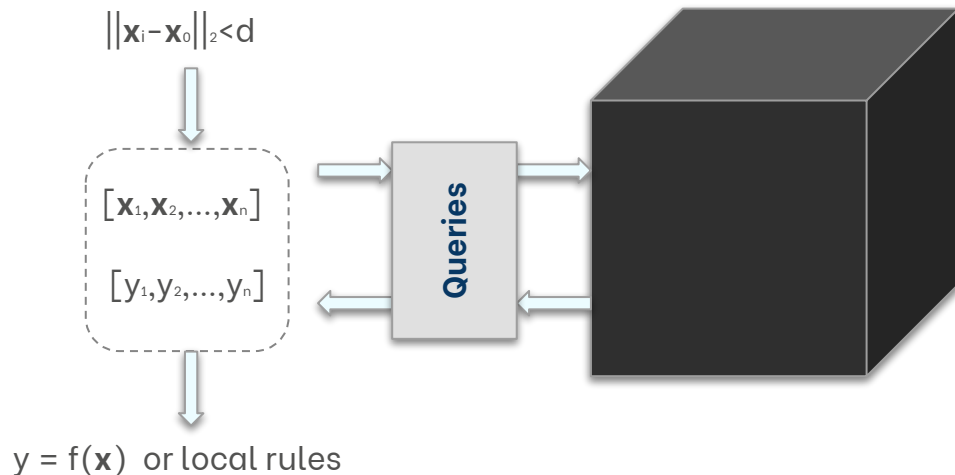


Metodi perturbativi

Agnostici rispetto al tipo di modello

Metodi disegnati per generare spiegazioni **locali** in maniera **agnostica**.

Spiegazioni generate ricostruendo un **modello semplificato** che approssimi il modello originale nelle vicinanze del punto da spiegare.

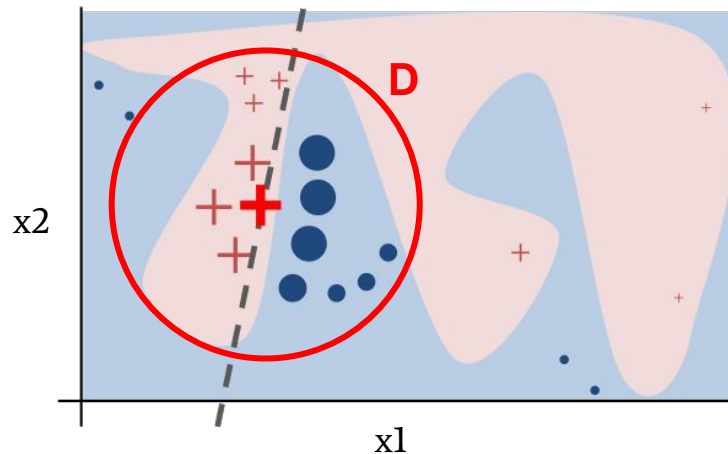


Local Interpretable Model-agnostic Explanations (LIME)

Rientra tra i metodi più popolari,
precursore approcci perturbativi.

Metodo:

- Genera **N** punti in un intorno **D** del punto da spiegare.
- Ottieni la risposta del modello per questo insieme di punti.
- Costruisci un classificatore lineare usando le x, y ottenute nei passi precedenti.

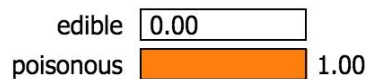


<https://arxiv.org/pdf/1602.04938.pdf>

LIME

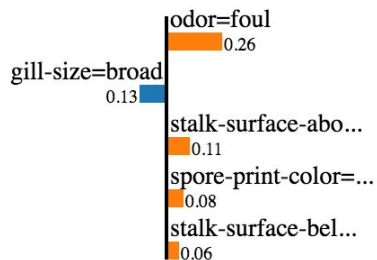
Esempio

Prediction probabilities



edible

poisonous



Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

- Spiegazioni generate usando diversi iperparametri (N punti, distanza D, etc)
- Può essere applicato a problemi su dati strutturati e non strutturati
- Spiegazioni possono non convergere.
- Spiegazioni non sono **prescrittive**.

Shapley values

Teoria dei giochi

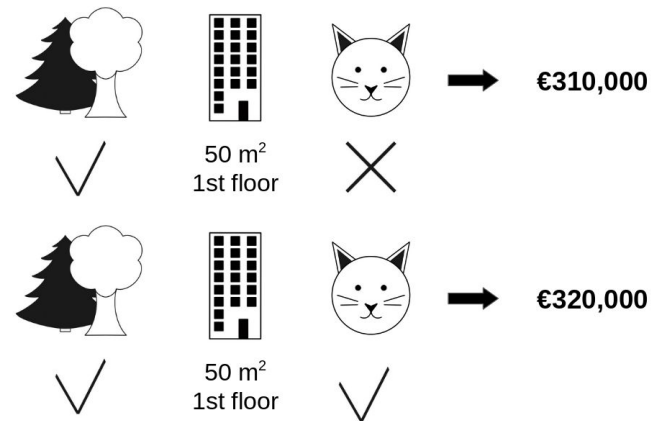
Concetto proveniente dalla teoria dei giochi:
come ridistribuire la ricompensa di un gioco a cui ha partecipato un gruppo di giocatori in maniera cooperativa?

Coefficienti di Shapley definiscono una maniera per distribuire ricompensa tra partecipanti.

Applicato al machine learning:

Giocatori → **Features**

Ricompensa → **Output del modello**



<https://christophm.github.io/interpretable-ml-book/>

Coalitions $\xrightarrow{h_x(z')}$ Feature values

Instance x

$$x' = \begin{array}{c|c|c} \text{Age} & \text{Weight} & \text{Color} \\ \hline 1 & 1 & 1 \end{array}$$

$$x = \begin{array}{c|c|c} \text{Age} & \text{Weight} & \text{Color} \\ \hline 0.5 & 20 & \text{Blue} \end{array}$$

Instance with
"absent"
features

$$z' = \begin{array}{c|c|c} \text{Age} & \text{Weight} & \text{Color} \\ \hline 1 & 0 & 0 \end{array}$$

$$z = \begin{array}{c|c|c} \text{Age} & \text{Weight} & \text{Color} \\ \hline 0.5 & \cancel{20} & \cancel{\text{Blue}} \\ & \downarrow & \downarrow \\ & 17 & \text{Pink} \end{array}$$

<https://christophm.github.io/interpretable-ml-book/>

SHAP

Approssimazione Shapley Values

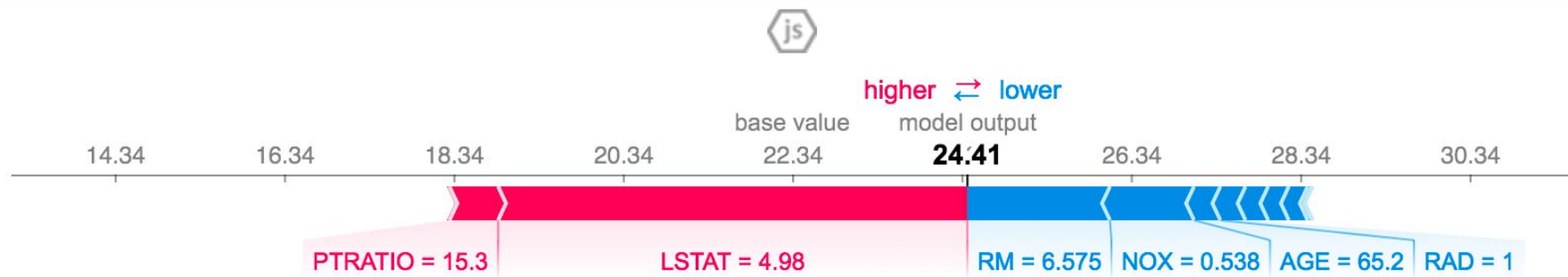
Metodo definisce una strategia per ottenere un'approssimazione dei valori di Shapley usando un approccio molto simile a LIME.

KernelSHAP → Metodo completamente agnostico rispetto al modello

TreeSHAP → Implementazione per modelli basati su alberi decisionali, molto più veloce

SHAP

Approssimazione Shapley Values



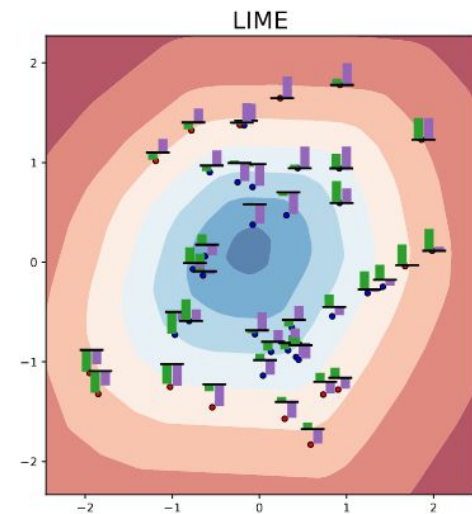
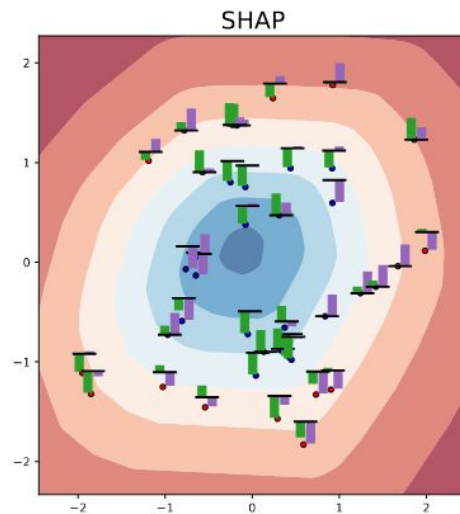
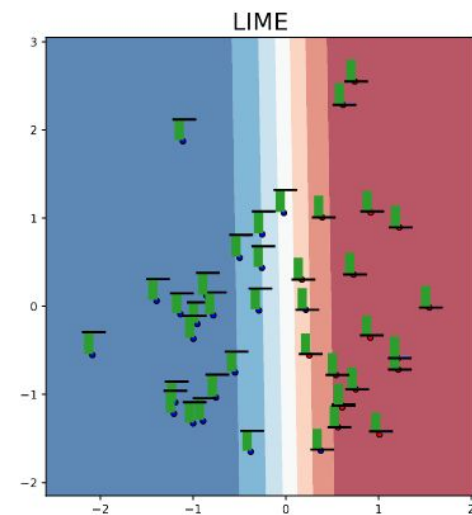
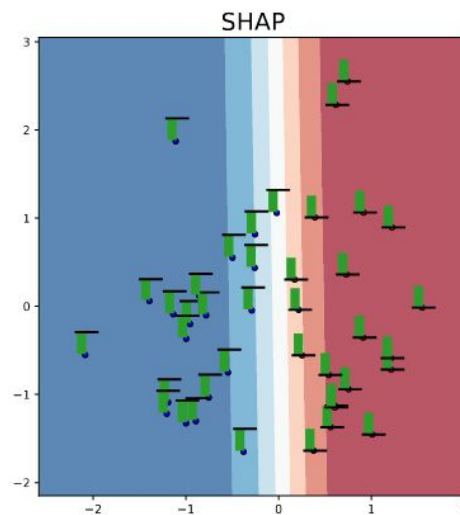
- Libreria SHAP offre oltre al metodo stesso librerie di visualizzazione molto curate
- KernelSHAP e' decisamente lento.
- Così come per LIME la generazione di punti sintetici non tiene conto delle dipendenze statistiche tra features.

Robustezza spiegazioni

Per alcuni tipi di problemi, librerie perturbative possono essere affette da problemi di robustezza

→ Piccolo cambiamento input può **cambiare completamente spiegazione.**

Importante monitorare qualità spiegazioni.



Esercizio SHAP

https://github.com/Clearbox-AI/Corso_MLOps/blob/main/sessione3/SHAP_examples.ipynb

Spiegazioni tramite esempi

Saper indicare quali dati di allenamento hanno influito su una particolare predizione.



Spiegazioni tramite esempi

Saper indicare quali dati di allenamento hanno influito su una particolare predizione.

ID immobile	locali	tipo	Superficie (m2)	bagni	giardino	Venduto a
n	3	villa	150	2	si	315000 €

Spiegazioni tramite esempi

Saper indicare quali dati di allenamento hanno influito su una particolare predizione.

ID immobile	locali	tipo	Superficie (m2)	bagni	giardino	Venduto a
n	3	villa	150	2	si	315000 €

Questo immobile proveniente dai dati di allenamento e' molto simile:

ID immobile	locali	tipo	Superficie (m2)	bagni	giardino	Venduto a
424	3	villa	140	2	si	310000 €

Spiegazioni tramite esempi

Saper indicare quali dati di allenamento hanno influito su una particolare predizione.

Problema: come definire similitudini tra punti?

Approccio più immediato → Nearest Neighbours utilizzando input stessi o rappresentazioni intermedie interne ai modelli (es: attivazioni in reti neurali)



https://beenkim.github.io/papers/KIM2016NIPS_MMD.pdf

Esempi controfattuali

ragionamento tramite scenari
ipotetici

Esempi che **cambiano la predizione** in una determinata direzione stando all'interno di **vincoli** specifici.

Es: *Se avessi avuto 5 anni in più il prestito sarebbe stato accettato.*

Impostati come problema di ottimizzazione vincolata nello spazio delle features.

$$\|\mathbf{x} - \mathbf{x}'\| \leq \delta$$



$$|f(\mathbf{x}) - f(\mathbf{x}')| > \epsilon$$

Metodi intrusivi

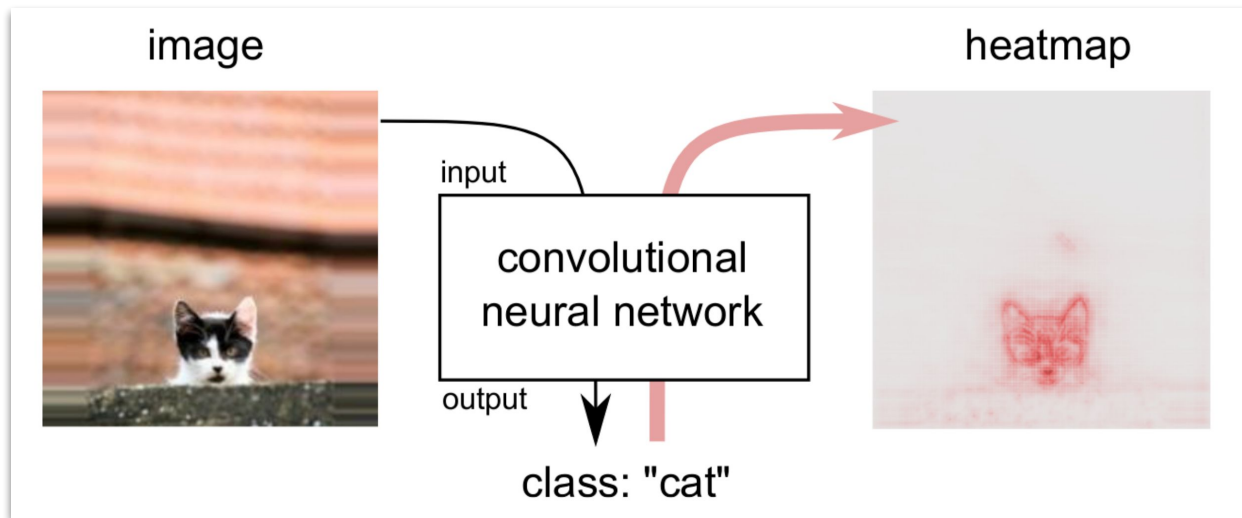
Aprire la black-box



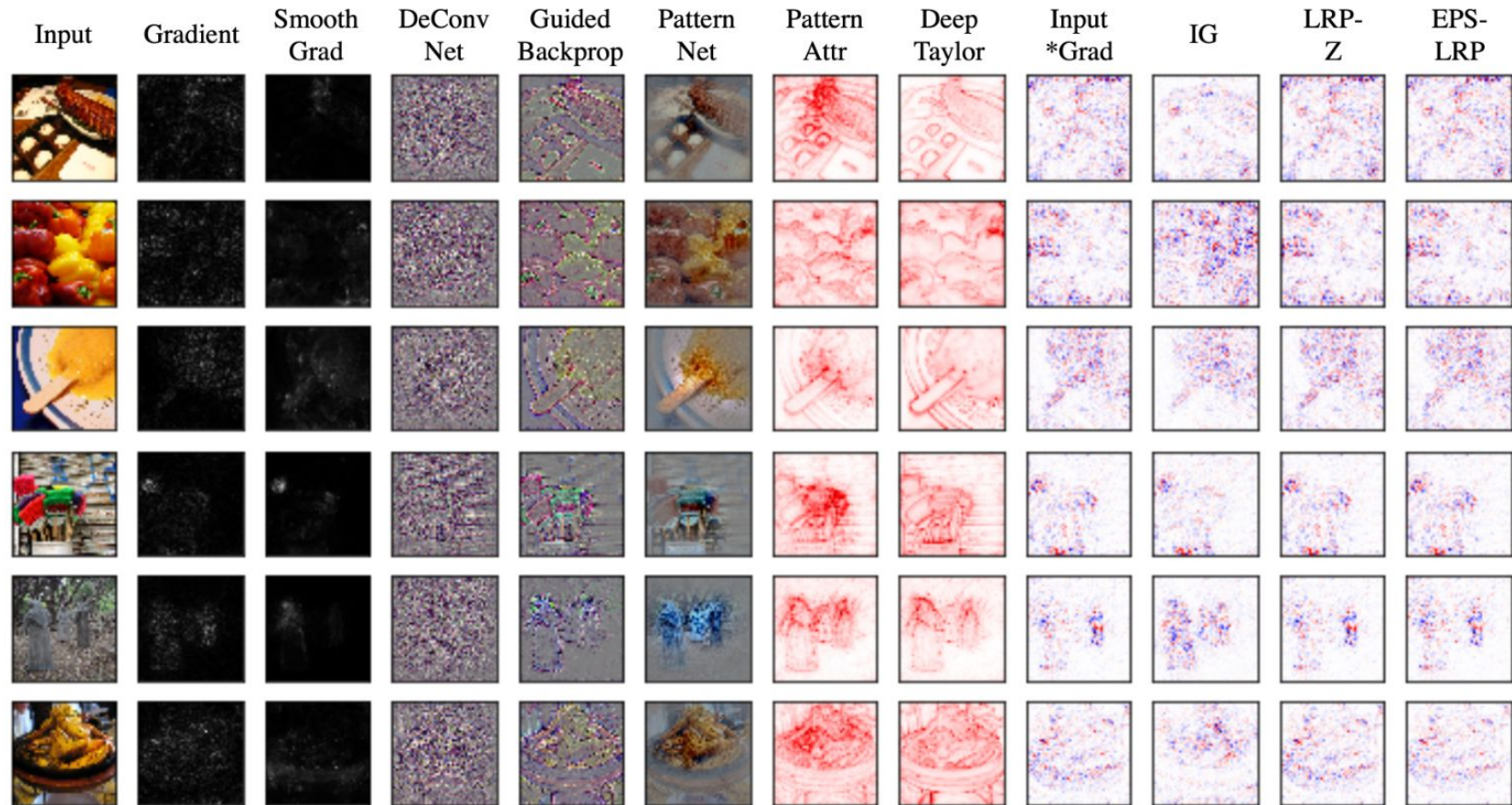
Metodi intrusivi

Aprire la black-box

Esempio: Deep Taylor Decomposition



Source: Montavon et al. (ICML 2016)



Singh A, Sengupta S, Lakshminarayanan V. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging*. 2020; 6(6):52. <https://doi.org/10.3390/jimaging6060052>

Librerie interpretazione reti neurali

CAPTUM (<https://github.com/pytorch/captum>). Librerie interpretabilità per modelli scritti in PyTorch.

iNNvestigate (<https://github.com/albermax/innvestigate>). Stesso ma per modelli scritti in Keras-Tensorflow (solamente 1.0, 2.0 in fase di sviluppo)

Clustering spiegazioni

Passare da analisi locale
spiegazioni ad analisi globale.
Come?

Raggruppando le spiegazioni in
maniera non supervisionata e
analizzando i clusters trovati
come **comportamenti** del
modello.

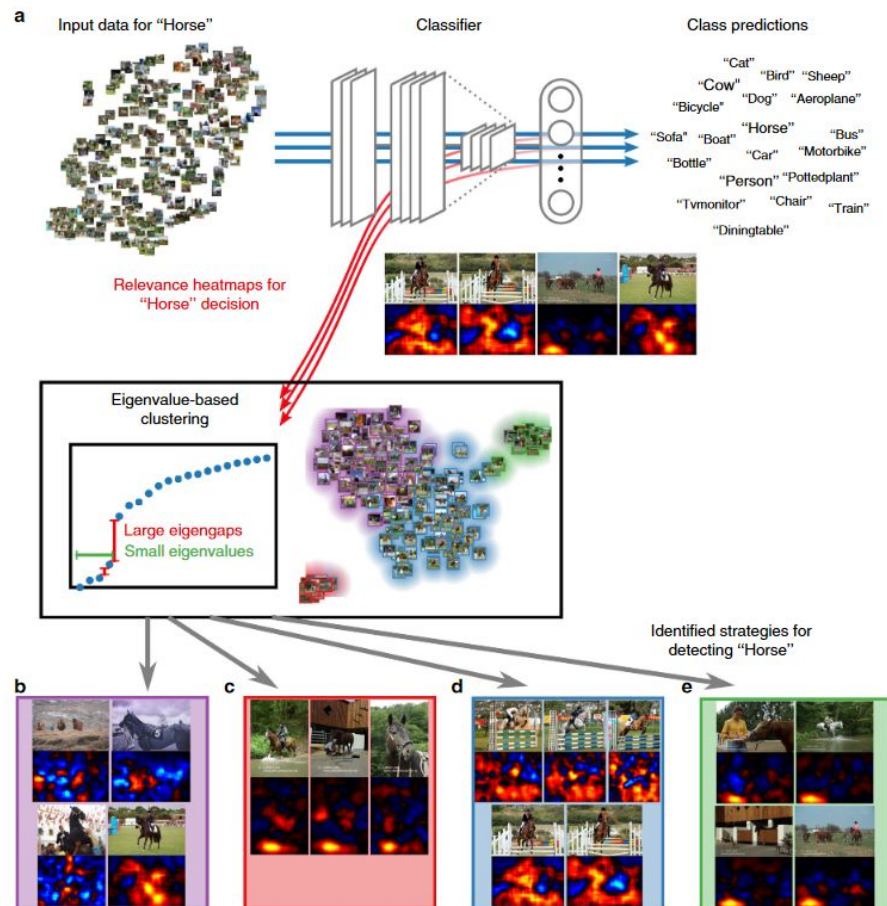
Idealmente approccio
permette di isolare
comportamenti non desiderati.

Clustering spiegazioni

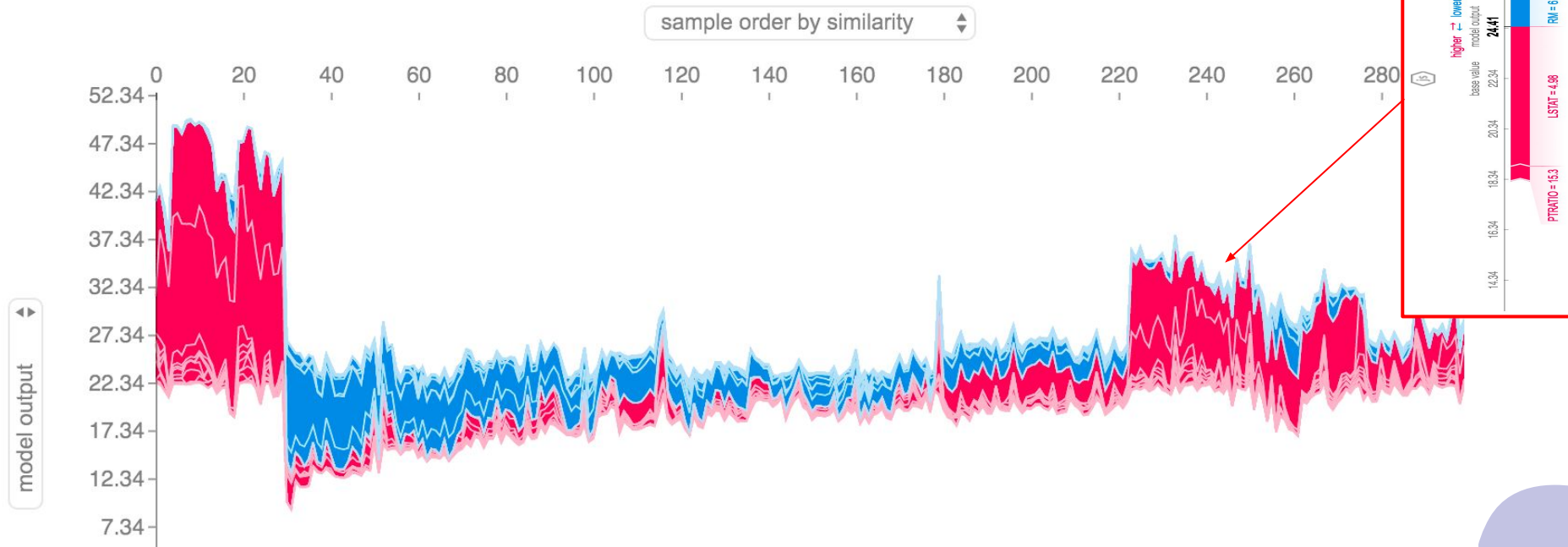
Passare da analisi locale spiegazioni ad analisi globale. Come?

Raggruppando le spiegazioni in maniera non supervisionata e analizzando i clusters trovati come **comportamenti** del modello.

Idealmente approccio permette di isolare comportamenti non desiderati.



Clustering spiegazioni SHAP



Esercizio clustering con SHAP

https://github.com/Clearbox-AI/Corso_MLOps/blob/main/sessione3/SHAP_examples.ipynb

Analisi degli errori: motivazione

Un modello validato ed interpretabile sarà comunque soggetto ad errori.

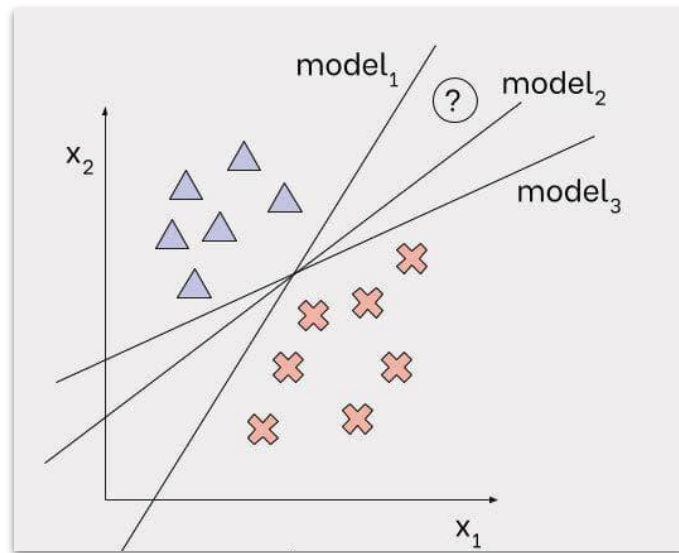
Dobbiamo essere in grado di stimare la **probabilità di incorrere** in errore e di permettere a modelli di optare di non dare una risposta.

Incertezza epistemica

o riducibile

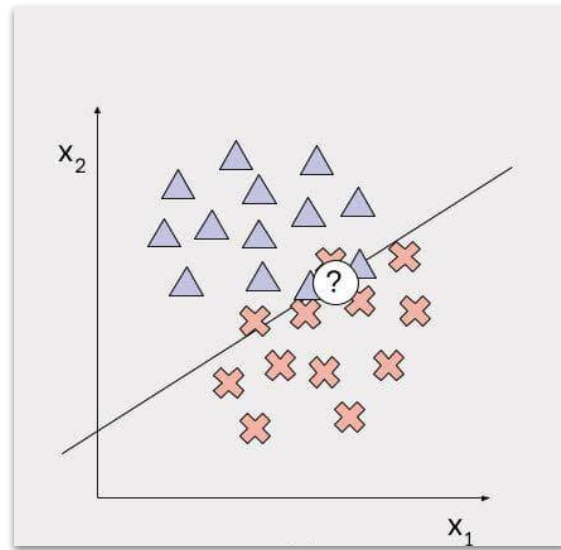
Incertezza epistemica rappresenta l'incertezza dovuta ad una **conoscenza incompleta** del problema analizzato.

In ambito machine learning e' solitamente associata a mancanza di informazione a livello di dati.



Incertezza aleatoria o irriducibile

Questo tipo di incertezza e' associata ad una presenza di **rumore** a livello di dati che non può essere ridotto tramite feature engineering o raccolta dati.



Calibrazione modelli

Definizione

Approccio comune in ambito machine learning: incorporare le varie sorgenti di incertezza all'interno di un'unica quantità, la **confidenza del modello**.

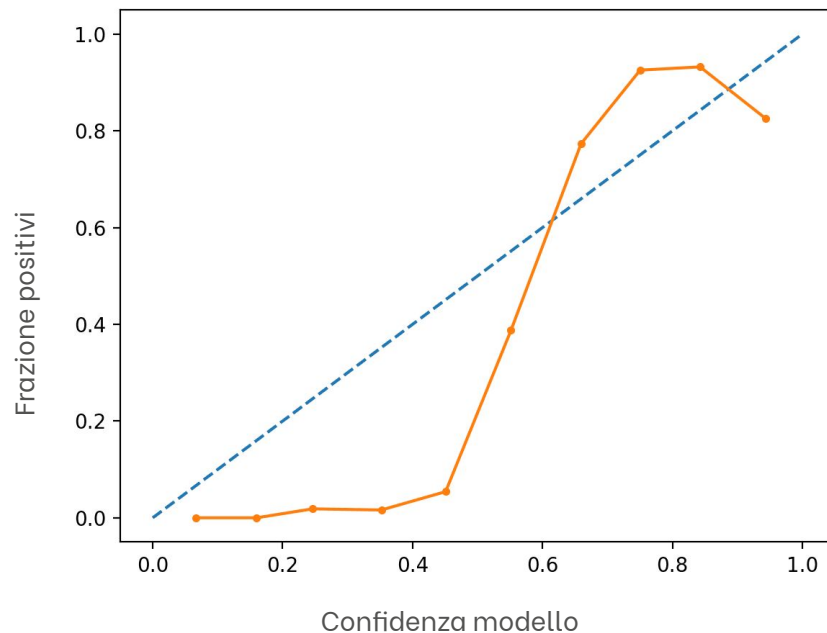
Problema sorge quando in presenza di modelli cosiddetti non calibrati:

es. una confidenza del 90% in un'etichetta deve corrispondere dal vivo in un rateo di successo analogo.

Calibrazione modelli

Metodi

Calibrazione di un modello può essere quantificata tramite la curva di calibrazione (o reliability diagram)



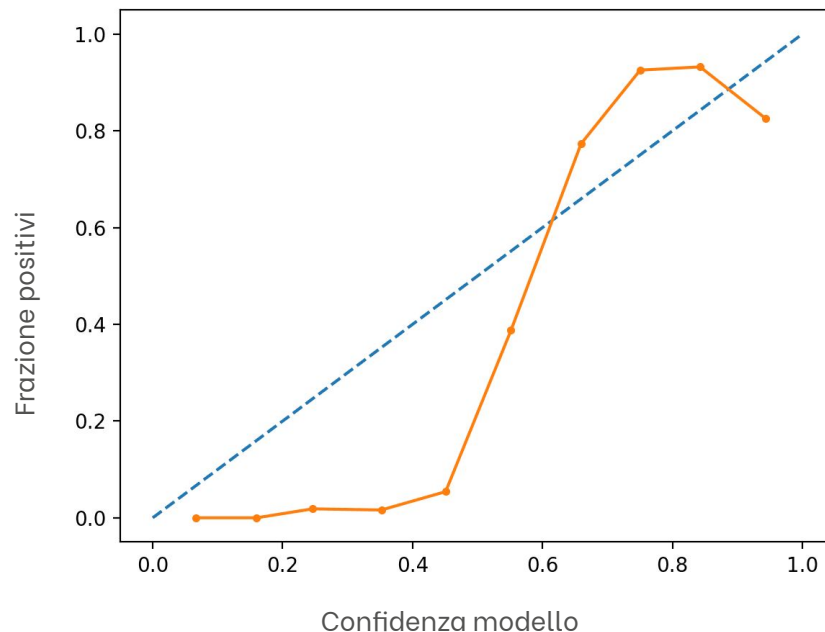
Calibrazione modelli

Metodi

Calibrazione di un modello può essere quantificata tramite la curva di calibrazione (o reliability diagram)

Metodi per migliorare calibrazione di un modello non calibrato:

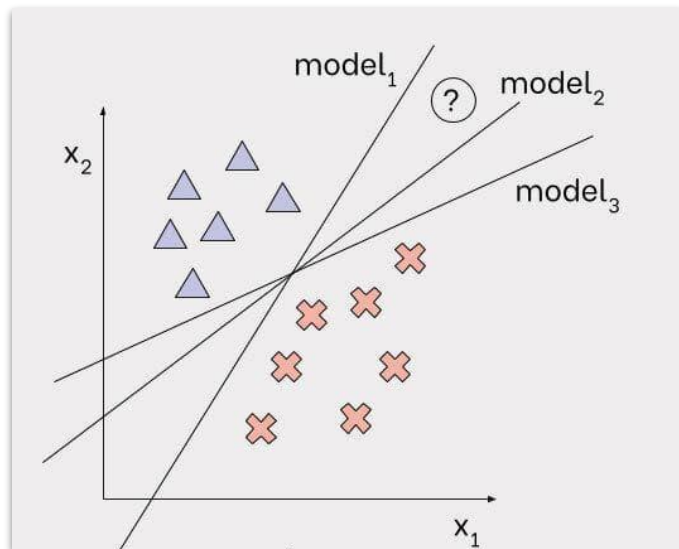
- **Platt scaling:** fit di una funzione di tipo sigmoide da aggiungere all'output del modello
- **Isotonic regression:** utilizzo di una funzione di regressione isotonica



Calibrazione modelli

Esercizio

https://github.com/Clearbox-AI/Corso_MLOps/blob/main/sessione3/Calibrazione_Anomaly_Detection.ipynb



Anomaly detection

Definizione limiti operativi modello

Prima di spostarsi nella fase di produzione bisogna definire un **range operativo** per un dato modello.

Modelli sono allenati su determinate distribuzioni di dati, devo assicurarmi che i dati in produzione provengano da distribuzioni simili, altrimenti il modello starebbe lavorando in un range sconosciuto.

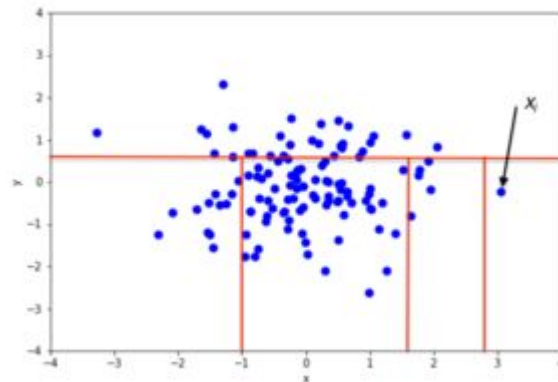
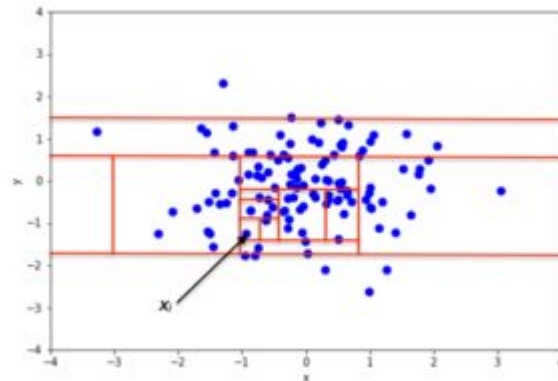
Come automatizzare questo processo di **rejection** di eventuali punti mai visti?
Tramite **anomaly detection**.

Isolation Forests

Trovare anomalie tramite alberi decisionali

Idea: quando creiamo un albero decisionale i punti più anomali sono i più facili da isolare tramite splitting. (Foglie contenenti anomalie tendono ad essere vicine alla radice dell'albero)

Un Isolation forest consiste in un numero di isolation trees dove i punti anomali sono quelli isolati usando il percorso più breve.



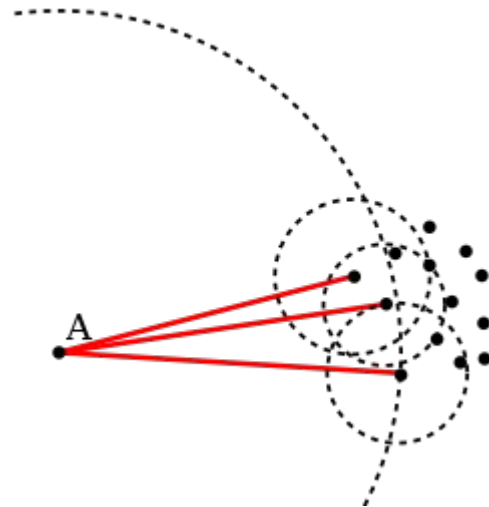
Local Outlier Factor

Analisi densità locali

Idea: Comparare la densità locale di un punto con densità locale di punti vicini.

Così come isolation forest e' basata su alberi decisionale, LOF e' basato su tecniche di tipo **Nearest Neighbours**.

Caratterizzato da stesse limitazioni associate a kNN: quale **metrica** usare per calcolare distanze tra punti?



Anomaly detection

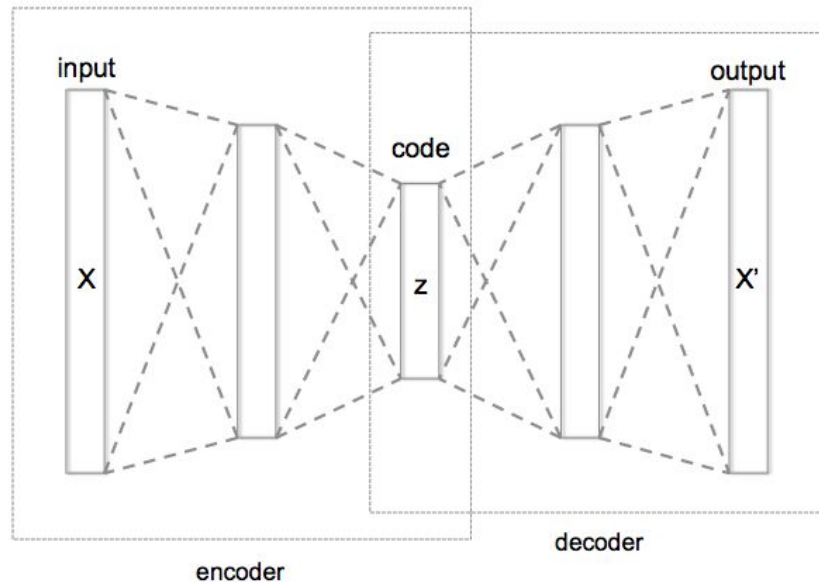
Esercizio con scikit-learn

https://github.com/Clearbox-AI/Corso_MLOps/blob/main/sessione3/Calibrazione_Anomaly_Detection.ipynb

AutoEncoders

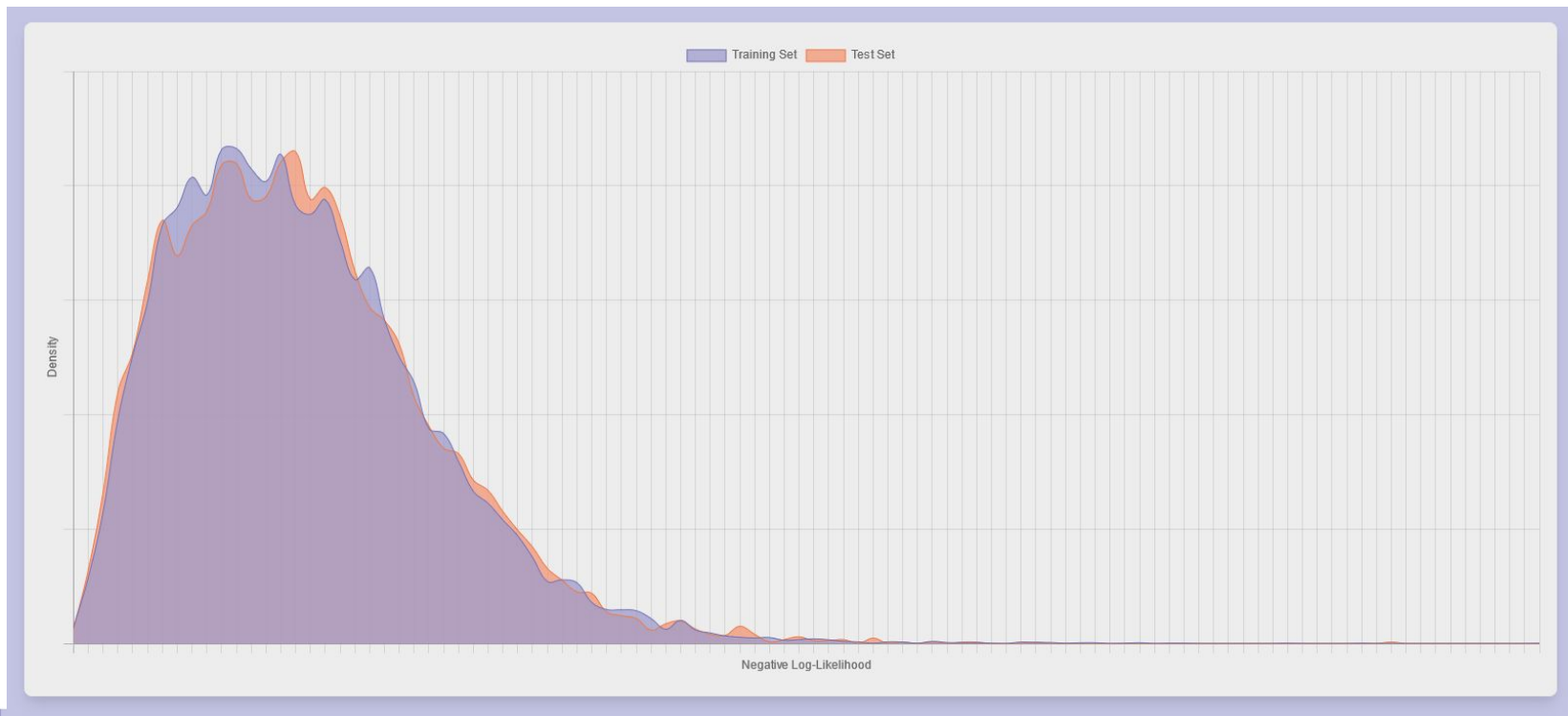
Idea: allenare un modello (basato su reti neurali) a **comprimere** e **ricostruire** i dati di allenamento del problema in esame.

Performando questo task su dati non visti prima l'errore di ricostruzione tenderà ad essere più alto per punti che non appartengono alla distribuzione di allenamento.



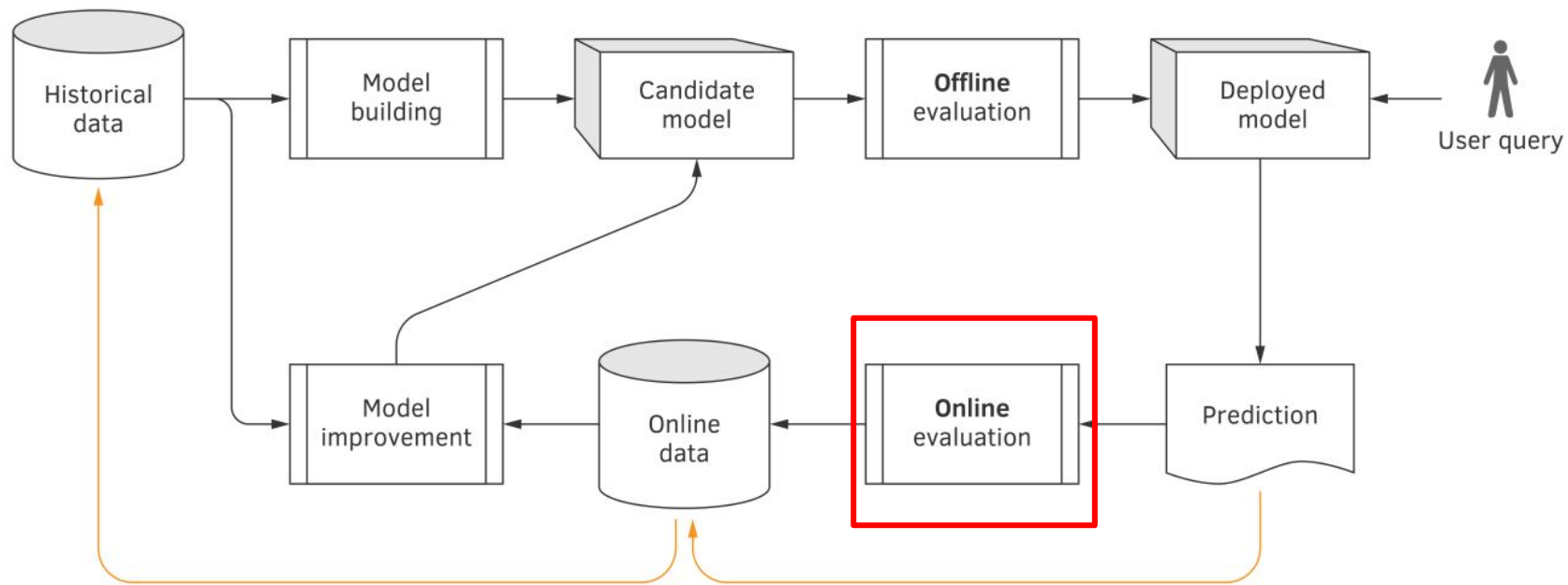
AutoEncoders

Esempio errore ricostruzione



Monitoraggio e miglioramento continuo

Prossima lezione





Thanks for Reading

Feel free to contact us:



www.clearbox.ai



shalini@clearbox.ai
giovannetti@clearbox.ai

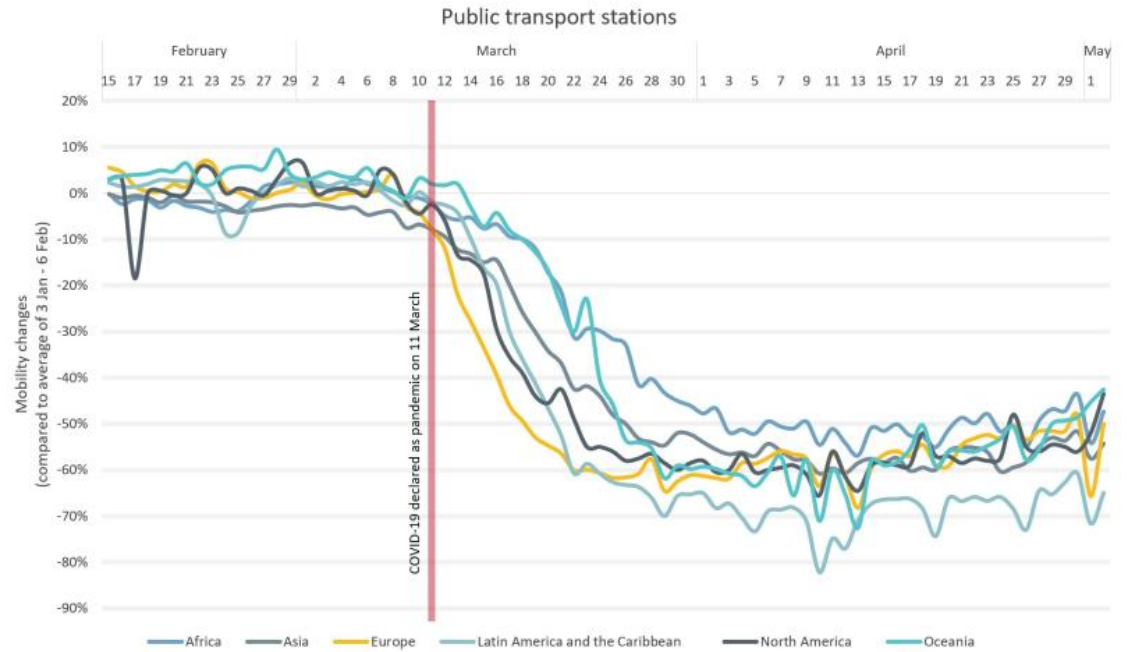


[@ClearboxAI](https://twitter.com/ClearboxAI)

Monitoraggio



Data-drift



Concept-drift

