To cite this presentation: Pendyala, V.S. (2024) "Explainability of AI/ML models for a socially sustainable AI growth". University of Bolton's November to Remember 2024 Inaugural Lecture

Explainability of AI/ML models for a socially sustainable AI growth

Vishnu S. Pendyala, Ph.D.

San Jose State University

Video Recording: https://www.youtube.com/watch?v=R8HR9Y3vG-8

Source: The New Yorker, November 13, 2023

PROFILES

WHY THE GODFATHER OF A.I. FEARS WHAT HE'S BUILT

Geoffrey Hinton has spent a lifetime teaching computers to learn. Now he worries that artificial brains are better than ours.

”We fear what we don't understand"

- spoken by the character of Dr. Jonathan Crane (aka Scarecrow) in the "Batman Begins" movie,

AI makes mistakes sometimes

- Decides unfairly
- Overfits causing privacy concerns
- Hallucinates giving incorrect answers
- And often, no one can accurately explain its behavior!

Pendyala, V., & Kim, H. (2024). Assessing the Reliability of Machine Learning Models Applied to the Mental Health Domain Using Explainable AI. *Electronics*, *13*(6), 1025.

"This work proves that merely achieving superlative evaluation metrics can be dangerously misleading and may infringe upon ethical horizons. A future direction is to investigate methods to quantify the effectiveness of machine learning models in terms of insights from their explainability."

Y. Liu, C. Tantithamthavorn, L. Li and Y. Liu, "Explainable AI for Android Malware Detection: Towards Understanding Why the Models Perform So Well?," 2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE), Charlotte, NC, USA, 2022, pp. 169-180,

"our results indicate that ML models classify malware based on temporal differences between malware and benign, rather than the actual malicious behaviors."

"We discover that temporal sample inconsistency in the training dataset brings over-optimistic classification performance (up to 99%F1 score and accuracy)."

Tian, Y., Ma, S., Wen, M. et al. To what extent do DNN-based image classification models make unreliable inferences?. Empir Software Eng 26, 84 (2021).

"we applied our approach to 18 pre-trained single-label image classification models and 3 multi-label classification models, and then examined their inferences on the ImageNet and COCO datasets. We found that unreliable inferences are pervasive. Specifically, for each model, ***more than thousands of correct classifications are actually made using irrelevant features***."

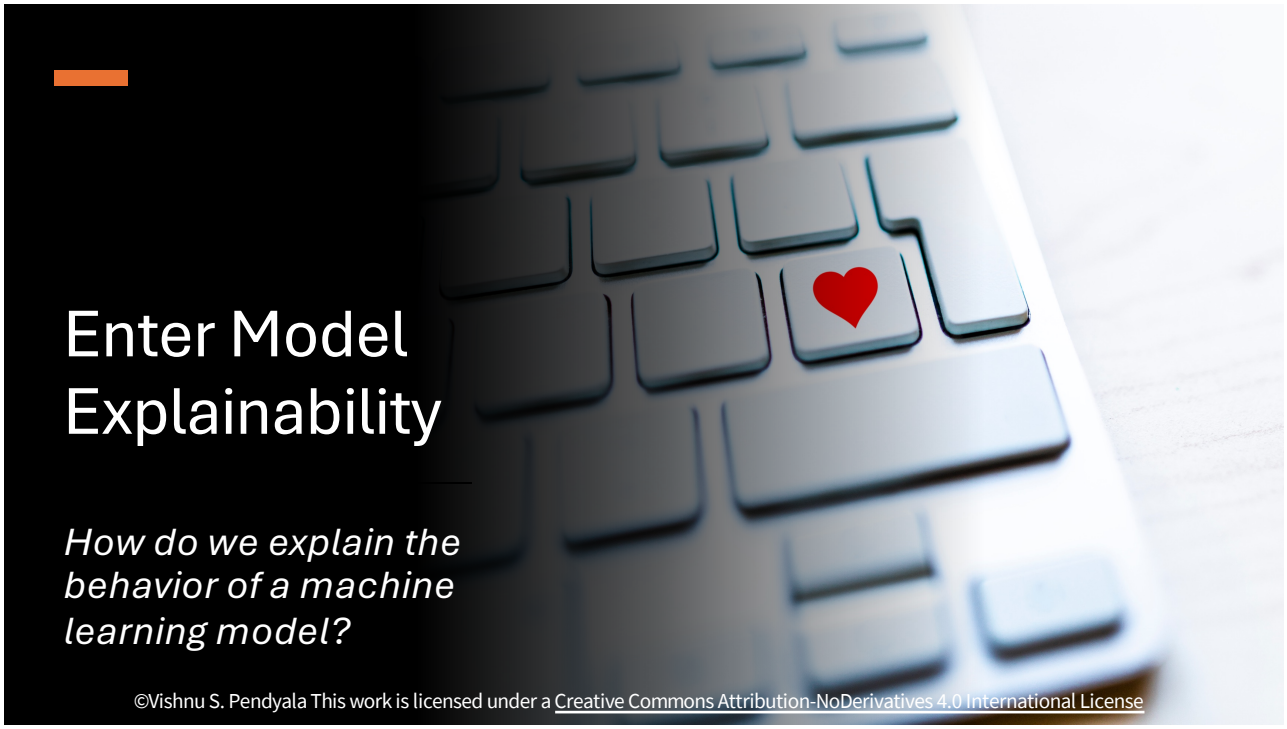
The "Right to explanation" in the GDPR

Recital 71

"…***to obtain an explanation*** of the decision reached after such assessment and to challenge the decision…"

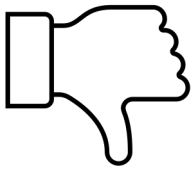
Articles 13, 14, and 15

"…of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, ***meaningful information about the logic involved***"



Enter Model Explainability

How do we explain the behavior of a machine learning model?



Explaining human behavior: Credit Denied!

- **Poor Credit Score**: Applicant had a history of late payments and defaults
- **Insufficient Income**: Applicant's income is too low relative to their existing debt and living expenses
- **Too Many Open Credit Accounts**: Applicant had multiple credit cards and loans - ability to manage additional credit is questionable
- ...and more

Machine Learning Model for Credit Approval: Features

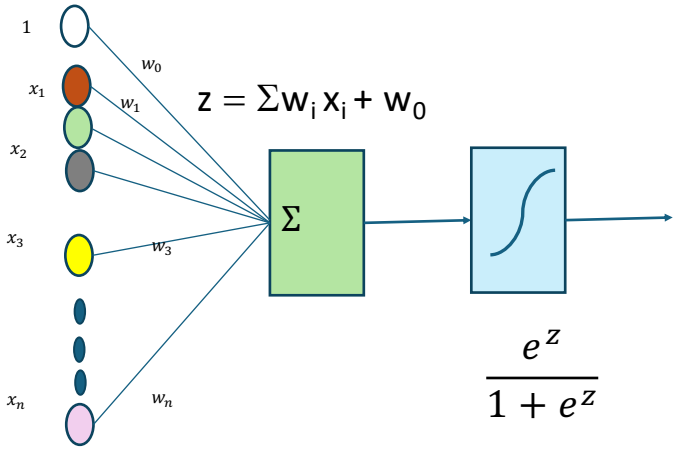
Gender	Age	Debt	Marital Status	Bank Customer
Education Level	Ethnicity	Years Employed	Prior Default	Employed
Credit Score	Late Payments	Citizen	Zip Code	Income

Target Variable: Approved

Source: UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/datasets/credit+approval

Logistic Regression model to approve or deny credit applications

$x_1 = credit\ score$
$x_2 = income$
$x_3 = prior\ default$
...
$x_i = debt$
$x_j = living\ expenses$
$x_k = years\ employed$
$x_n = education\ level$

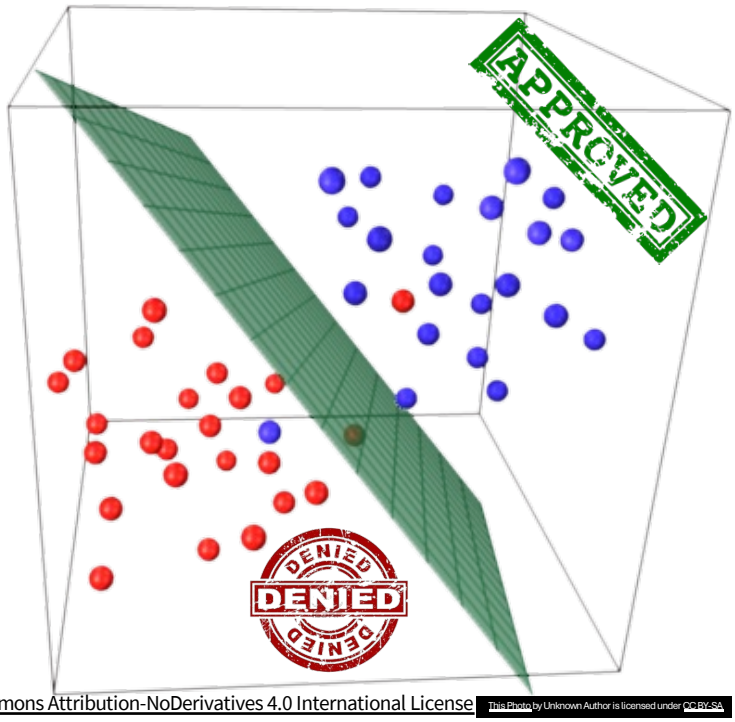


We can generate similar explanations as humans generate, based on the features and their weights

Linear Classification ML model for credit approval

$$y_i = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + \varepsilon$$

The w's indicate the importance of the features.

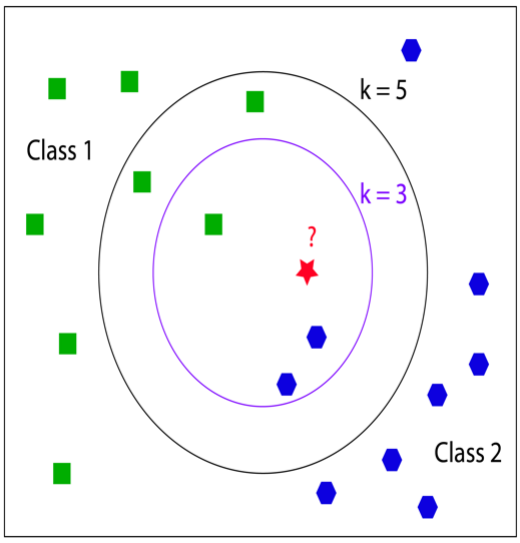


K Nearest Neighbors for Credit Approval

John's situation is similar to Tom, Dick, and Harry's

Tom's was approved but Dick's and Harry's was denied

We therefore denied John's application

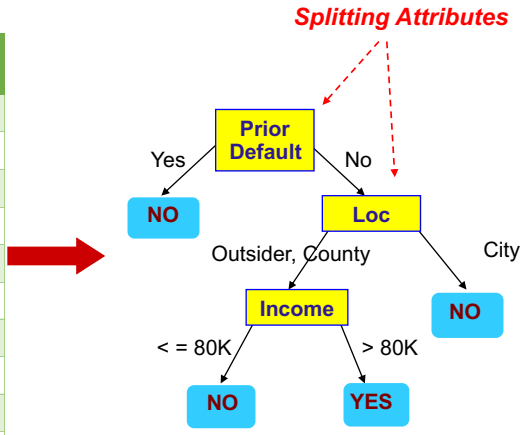


Approval using a Decision Tree Machine Learning Model

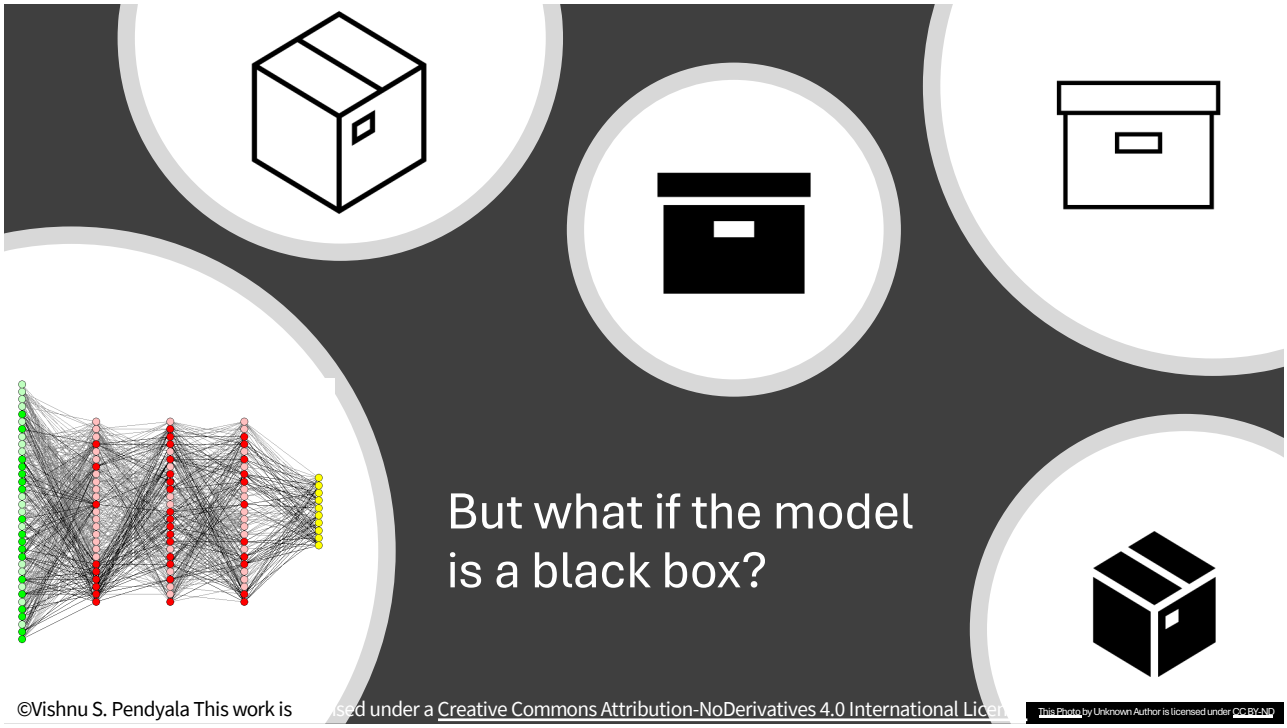
Prior Credit Approval Applications

Id	Prior Default	Locality	Income	Credit Approved ?
1	Yes	Outsider	125K	No
2	No	City	100K	No
3	No	Outsider	70K	No
4	Yes	City	120K	No
5	No	County	95K	Yes
6	No	City	60K	No
7	Yes	County	220K	No
8	No	Outsider	85K	Yes
9	No	City	75K	No
10	No	Outsider	90K	Yes

Training Data



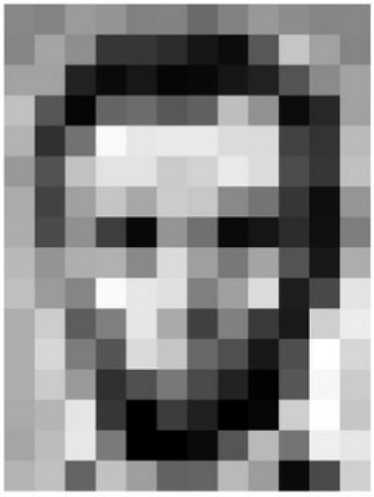
Model: Decision Tree

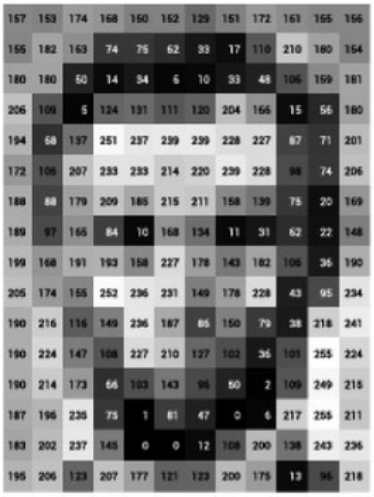


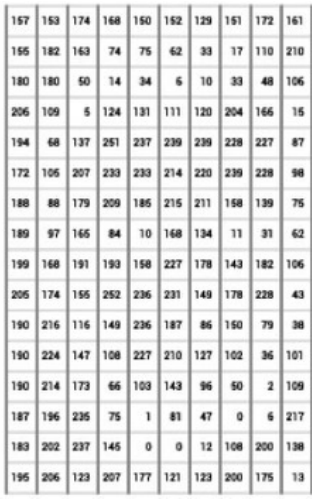
Can we create a model-agnostic method for explaining predictions of ML models?

- Explanations are not specific to the ML models – sometimes nearest neighbors or decision trees are not the best way to explain
- Easily replace or upgrade models without changing the explanation approach
- Explainability is now a layer on top of the black box ML model
- Easier benchmarking and comparisons of model interpretability

What if the features are not human-interpretable?



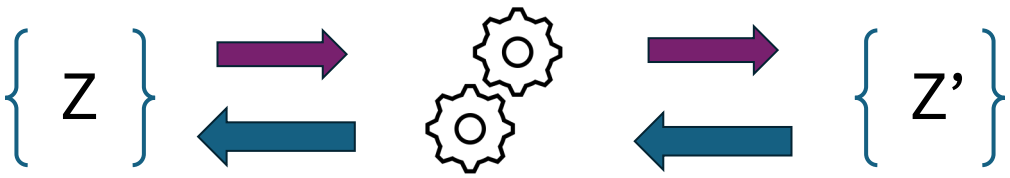




What if the features are not human-interpretable?

```
embedding = [
  0.234, -0.115, 0.587, 0.020, -0.367, 0.200, 0.150, 0.405, 0.100,
  -0.075, 0.289, 0.451, -0.201, 0.120, 0.340, 0.460, -0.069, 0.370,
  -0.199, 0.023, 0.135, -0.020, 0.250, 0.015, -0.233, 0.415, 0.056,
  0.089, -0.305, 0.670, 0.222, 0.300, -0.045, 0.090, 0.159, -0.028,
  0.016, 0.100, -0.220, 0.321, 0.067, -0.112, 0.245, 0.091, -0.030,
  0.375, 0.190, -0.177, 0.022, 0.060, 0.289, -0.098, 0.155, -0.074,
  0.118, 0.034, 0.205, -0.005, 0.019, 0.350, 0.088, -0.093, 0.417,
  0.204, -0.154, 0.136, -0.105, 0.209, 0.095, -0.050, 0.072, -0.189,
  # ... (remaining values)
  0.237, 0.152, 0.043, 0.092, -0.203, 0.155, 0.046, 0.360, -0.112,
  0.244, -0.087, 0.271, 0.048, 0.112, 0.190
]
```


Human interpretable knowledge representation



©Vishnu S. Pendyala This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License

Local Interpretable Model-agnostic Explanations (LIME)

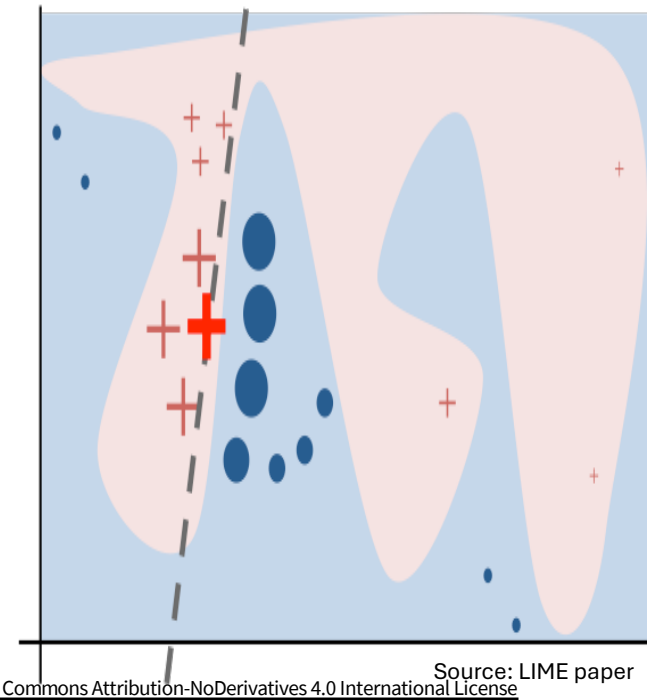
Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

- Provides local explanations for individual predictions (not global)
- Uses a surrogate interpretable model like the ones discussed earlier
- Local: Surrogate model is trained on data in a local context, typically around a specific data item
- Interpretable: Human-understandable even if the features are not
- Outputs feature importance scores for the prediction

	Local Interpretability	Global Interpretability
Focus	Explains individual predictions	Explains the overall model behavior
Scope	Specific instances or data points	Entire model or dataset
Methods	LIME, SHAP	Decision Trees, Rule-based models
Goal	Understand why a model made a specific prediction for a data item	Understand how the model works in general across all data
Application	Debugging, identifying biases, building trust	Model design, feature engineering, risk assessment
Use cases	Clinicians, loan officers needing explanations for specific decisions	Regulatory, where trust / compliance are crucial; scientific research
Level of Detail	High level of detail for specific predictions	Some level of detail sacrificed for comprehensiveness

How does LIME work?

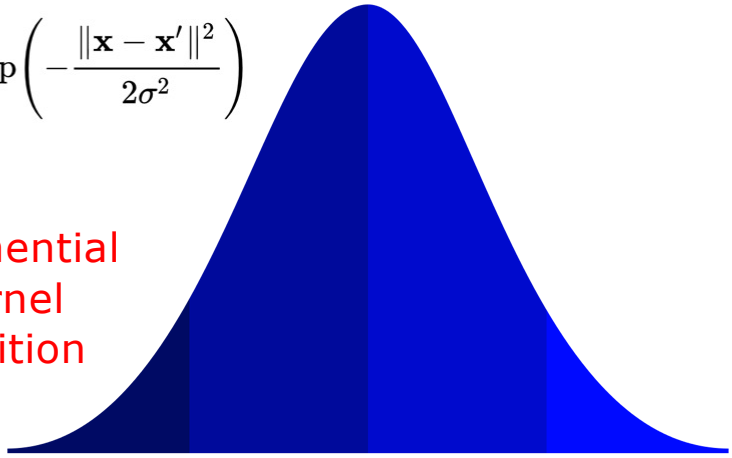
- Select an instance for explanation
- Generate perturbations around the instance
- Predict with the black-box model and use these labels
- Weigh the perturbations using a kernel, usually exponential
- Fit an interpretable model to the new data and generated labels
- Generate feature importance
- Visualize the explanation

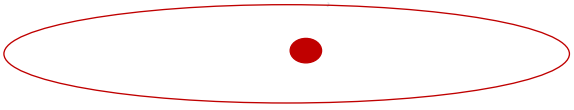


Source: LIME paper

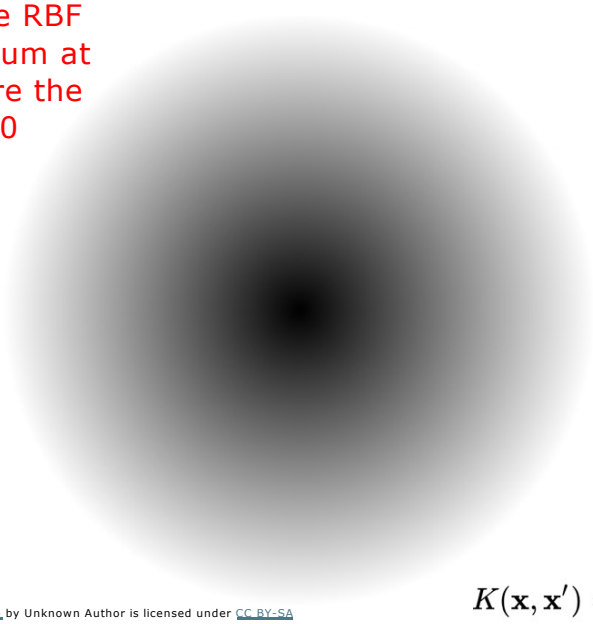
$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

Exponential Kernel Intuition





The value of the RBF Kernel is maximum at the center where the distance = 0
$e^0 = 1$



$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

LIME – the math

$$L(f, g, \pi_x) = \sum_{z \in Z} \pi_x(z)\left(f(z) - g(z')\right)^2, L = loss\ function$$

f *is the blackbox model,* g *is the surrogate interpretable model*

$$\pi_x(z) = \exp\left(\frac{-D(x,z)^2}{\sigma^2}\right); z \in \mathbb{R}^d \textit{is a sample in the perturbed set,} Z$$

$f(z)$ *is the label generated by the blackbox model*

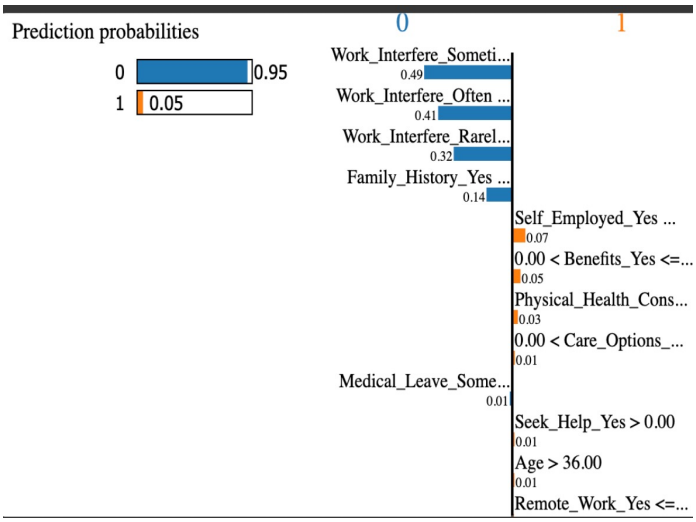
z' *is the interpretable version of* z; $g(z')$ *is the label from* g

$$\xi(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g)\ \textit{is the generated explanation}$$

G *is the set of interpretable models,* Ω *is the regularization*

Pendyala, Vishnu, and Hyungkyun Kim. "Assessing the Reliability of Machine Learning Models Applied to the Mental Health Domain Using Explainable AI." *Electronics* 13.6 (2024): 1025.

Feature	Value
Work_Interfere_Sometimes	0.00
Work_Interfere_Often	0.00
Work_Interfere_Rarely	0.00
Family_History_Yes	0.00
Self_Employed_Yes	0.00
Benefits_Yes	1.00
Physical_Health_Consequence_Yes	0.00
Care_Options_Yes	1.00
Medical_Leave_Somewhat easy	0.00
Seek_Help_Yes	1.00
Age	40.00
Remote_Work_Yes	0.00
Coworkers_Reach_Yes	0.00



Explaining Object Detection

SHapley Additive exPlanations (SHAP)

Scott, M., & Su-In, L. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30, 4765-4774.	Based on cooperative game theory, specifically using Shapley values; treats each feature as a player in a game
	Can be used for explanations at both global and local levels
	The model prediction is the sum of the SHAP values for each feature
	Calculates the average contribution of each feature to the prediction by considering all possible subsets of feature

Lloyd Shapley's Shapley Values

- Shapley, Lloyd S. "A value for n-person games." Contribution to the Theory of Games 2 (1953)
- Goal: Distribute total gains among players in a coalition based on their contributions
- Each player gets their fair share of the total gains based on their marginal contributions
- Considers all possible coalitions of players and their interactions
- Order independent, proportional weighting - fairness is key

Feature Subsets

Full set of features, F: **Credit Score (CS)**, **Annual Income (AI)**, **Loan Amount (LA)**, **Debt-to-Income Ratio (DIR)**, and **Employment Length (EL)**.

Feature Subset	Features Included
{LA}	Loan Amount
…	…
{EL}	Employment Length
{CS, AI}	Credit Score, Annual Income
…	…
{AI, EL}	Annual Income, Employment Length
{LA, DIR}	Loan Amount, Debt-to-Income Ratio
{CS, AI, EL}	Credit Score, Annual Income, Employment Length
…	…
{CS, LA, EL}	Credit Score, Loan Amount, Employment Length
{CS, DIR, EL}	Credit Score, Debt-to-Income Ratio, Employment Length
{CS, AI, DIR, EL}	Credit Score, Annual Income, Debt-to-Income Ratio, Employment Length
…	…
{AI, LA, DIR, EL}	Annual Income, Loan Amount, Debt-to-Income Ratio, Employment Length
{CS, AI, LA, DIR, EL}	Credit Score, Annual Income, Loan Amount, Debt-to-Income Ratio, Employment Length

SHAP – the math

$$\phi_i(f,x) = \sum_{S \sqsubseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right]$$

$\phi_i = Feature\ attribution\ for\ feature\ 'i'; x: Specific\ data\ item;$

$F: Set\ of\ all\ features; S: Feature\ subset; f: Blackbox\ model$

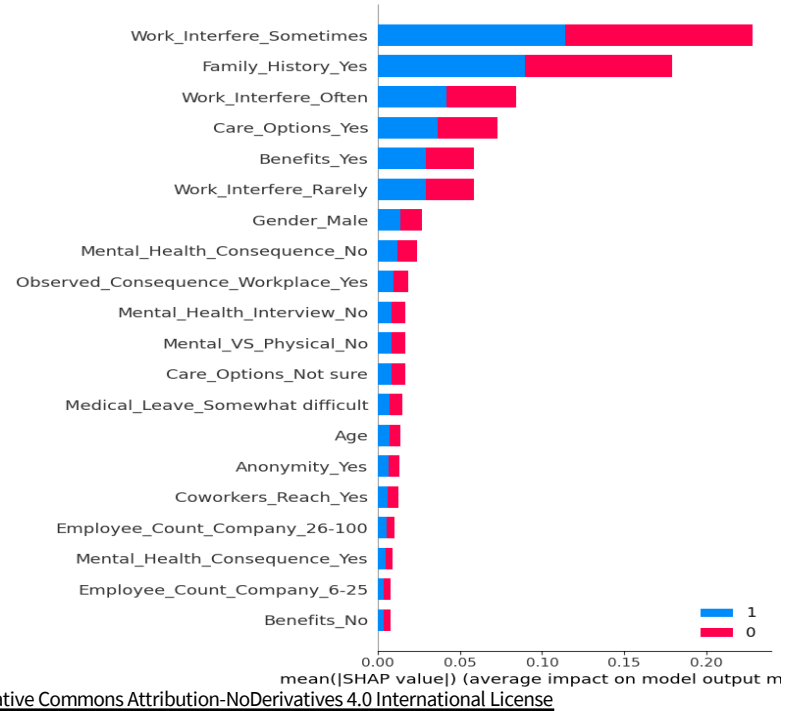
$\left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right] is\ the\ contribution\ of\ feature\ i\ in\ subset\ S$

$\frac{|S|!(|F|-|S|-1)!}{|F|!}$ is the weighting factor

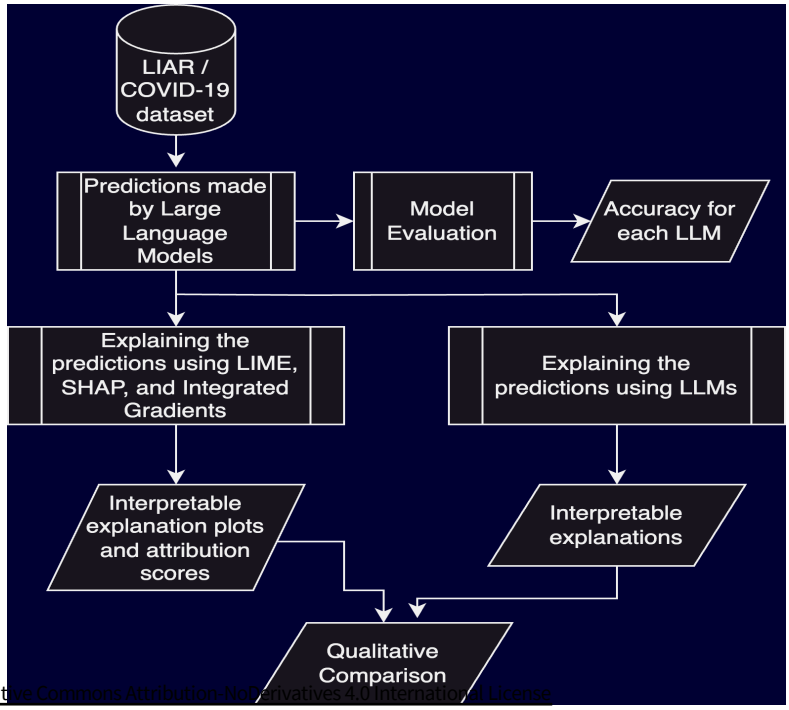
Features are "removed" by randomizing them

SHAP for mental health prediction model explainability

Pendyala, Vishnu, and Hyungkyun Kim. "Assessing the Reliability of Machine Learning Models Applied to the Mental Health Domain Using Explainable AI." Electronics 13.6 (2024): 1025.



Pendyala, Vishnu S., and Christopher E. Hall. 2024. "Explaining Misinformation Detection Using Large Language Models" *Electronics* 13, no. 9: 1673.

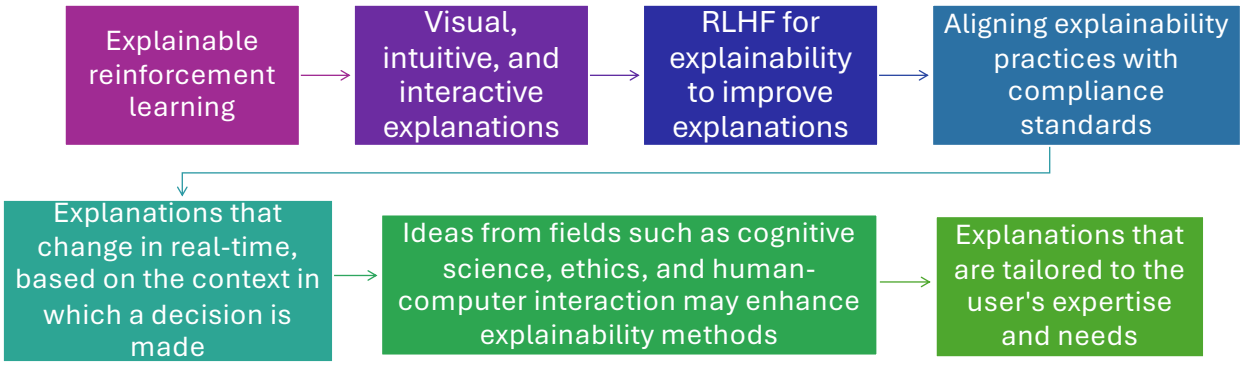


Explainable Misinformation Classification using Large Language Models

Legend: Negative Neutral Positive

Predicted Label	Experiment Type	Word Importance
Barely True (0.65)	IG	#s Please select the option that most closely describes the following claim by Donald Trump : H ill ary Cl inton has spoken such lies about my foreign policy . They said I want Japan to get nuclear weapons . Give me a break . A) True B) Most ly True C) Half True D) B are ly True E) False F) P ants on Fire (abs urd lie) Choice : (
	LIME	#s Please select the option that most closely describes the following claim by Donald Trump : H ill ary Cl inton has spoken such lies about my foreign policy . They said I want Japan to get nuclear weapons . Give me a break . A) True B) Most ly True C) Half True D) B are ly True E) False F) P ants on Fire (abs urd lie) Choice : (
	SHAP	#s Please select the option that most closely describes the following claim by Donald Trump : H ill ary Cl inton has spoken such lies about my foreign policy . They said I want Japan to get nuclear weapons . Give me a break . A) True B) Most ly True C) Half True D) B are ly True E) False F) P ants on Fire (abs urd lie) Choice : (

Future Directions



Questions?

and answers