Making Interpretable Neural Networks More Explainable

Krzysztof Nalborski

Head of Global AI/ML Pre-Sales – AI Innovation & Development (AIID)

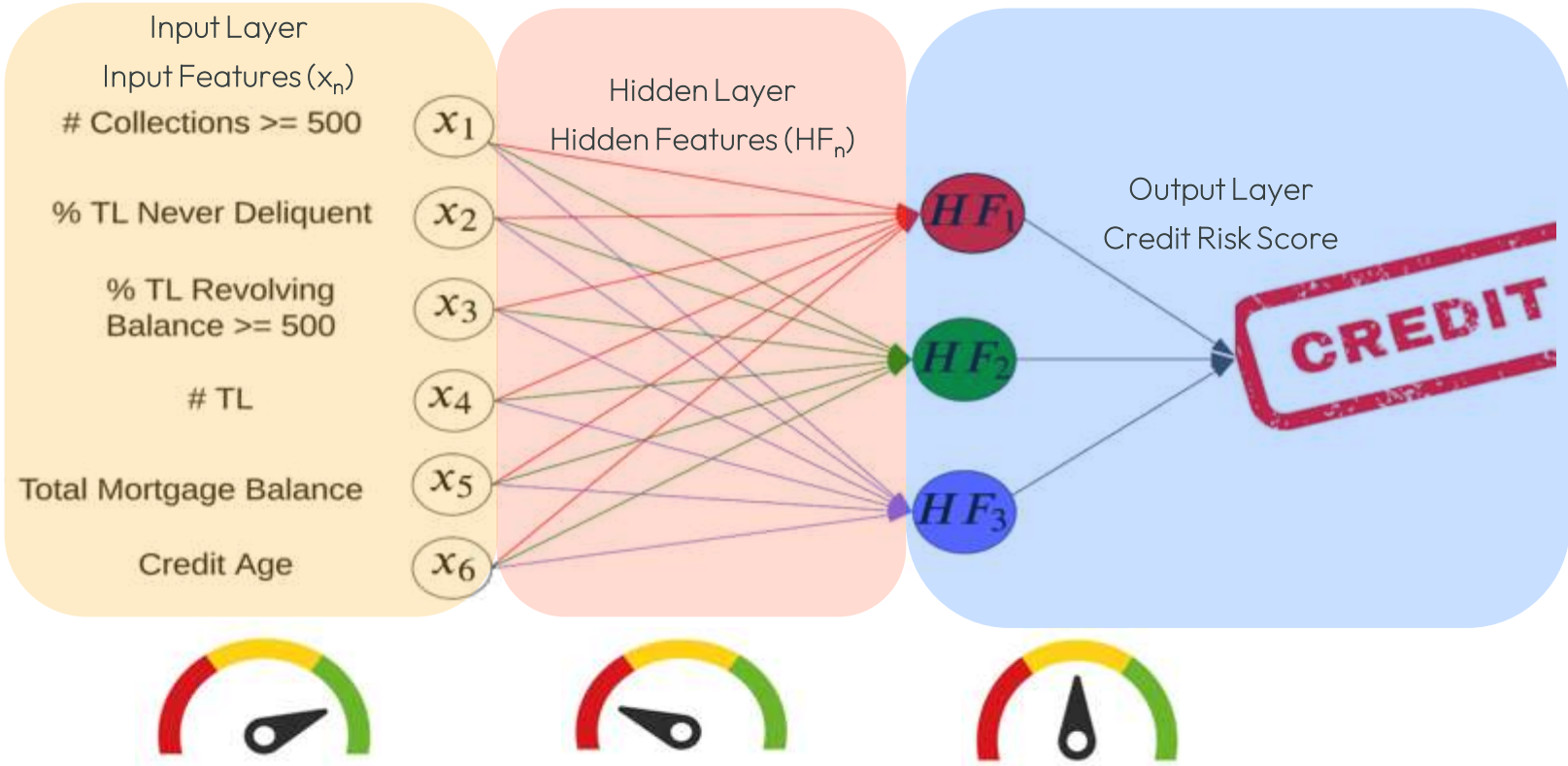
Research Collaborators:
Scott Zoldi – Chief Analytics Officer, FICO, AIID
Matt Kennel – Senior Principal Scientist, FICO, AIID
Mike Thompson – Senior Director, FICO, AIID

1. Neural networks are highly predictive but challenging to interpret and explain.

2. Hidden layer(s) is a key predictive component of a neural network, but fully interpreting and explaining its properties and typically dense connections is very challenging.

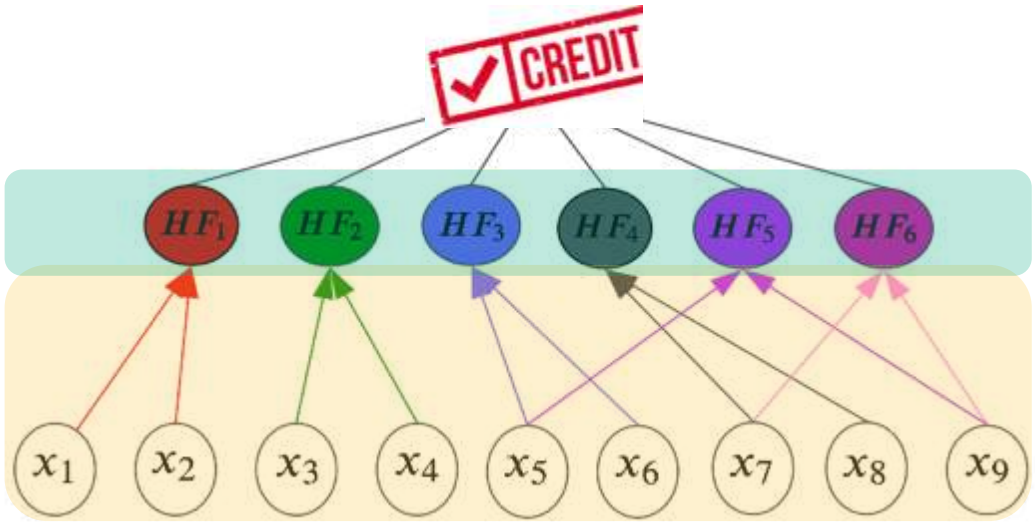
3. FICO's innovation addresses the interpretability and explainability issues with a novel neural network training method and construction of human traceable and readable reasons associated with each hidden feature.

Neural Networks: highly predictive but very challenging to interpret and explain

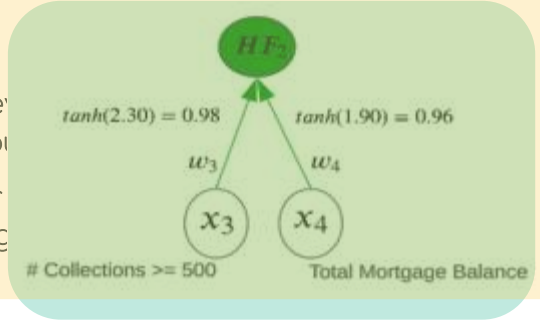


Interpretability and Explainability

Neural Networks: interpretability vs. explainability



- can be achie
  identify 2-inpu
- 2-input (or
  same training



-aining process to
ık in the appendix)
by found in the
constraints to the

Interpretability:

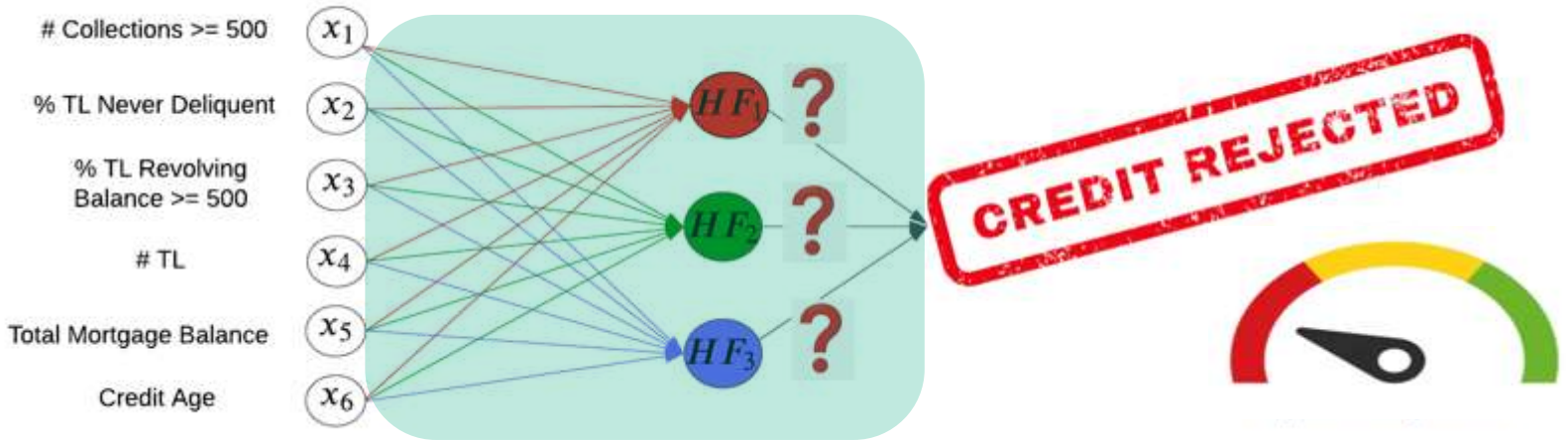
architectural transparency and sparsity where only a limited number of inputs (e.g., maximum of 2) can connect into each hidden feature

vs

Explainability:

deterministic set of human traceable and readable reasons that can explain the meaning of each hidden feature

Regulatory Requirements

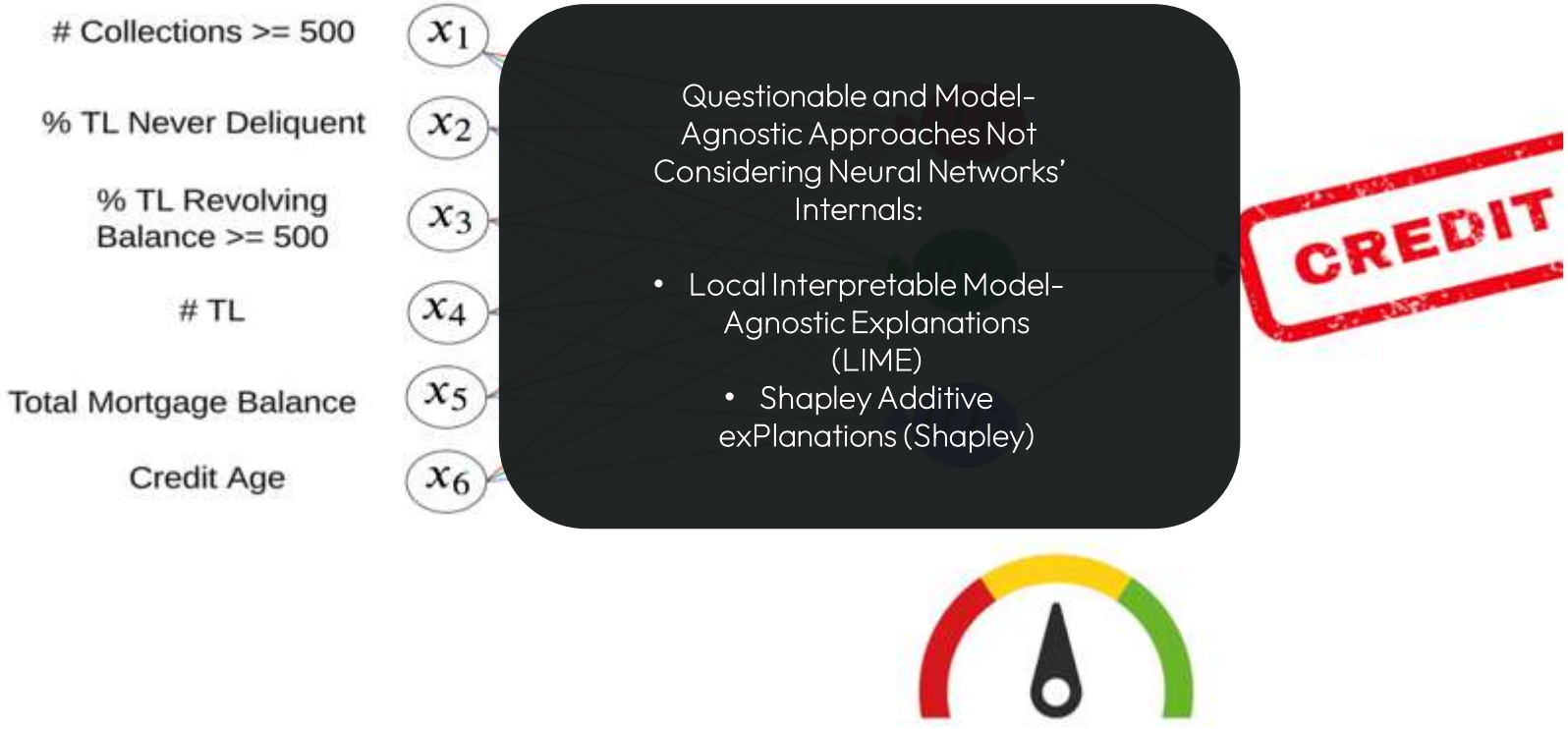


- Equal Credit Opportunity Act in the United States and General Data Protection Regulation (GDPR) in Europe require creditors to provide applicants who are denied credit with explanations regarding their rejected application.

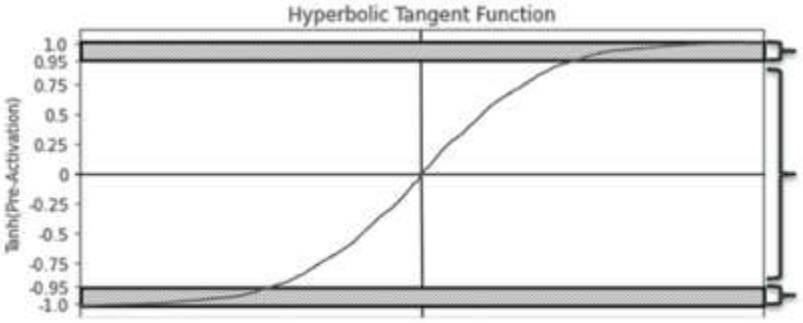
- "We regret to inform you that our AI system rejected your application" … will not be considered as a valid explanation.

- For credit risk decisions and to shift from the use of scorecards to neural networks, we need to be able to understand hidden features and enable less complexity in explanation by providing no more than 3-4 reasons per score.

Explainable AI: current approaches, challenges and motivation for our research



Hidden Layer: combinatorial explosion of reasons for a hidden feature value



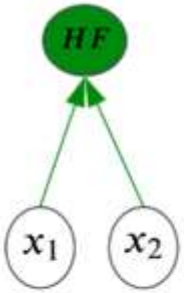
Positive Activation (1)
tanh(pre-activation term) >= 0.95

Non-activation (0)

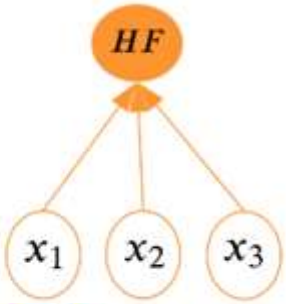
Negative Activation (-1)
tanh(pre-activation term) <= -0.95

- We define 3 tanh regions, as activation states, to minimize complexity of explanation
- pre-activation = feature value x weight

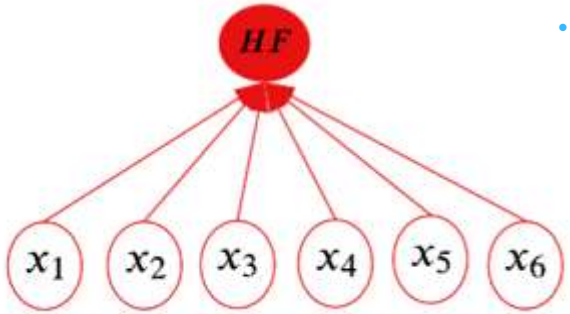
- We need to codify and assign human-traceable reasons for each activation state



3^2 possible activation states and reasons for the value of HF



3^3 possible activation states and reasons for the value of HF

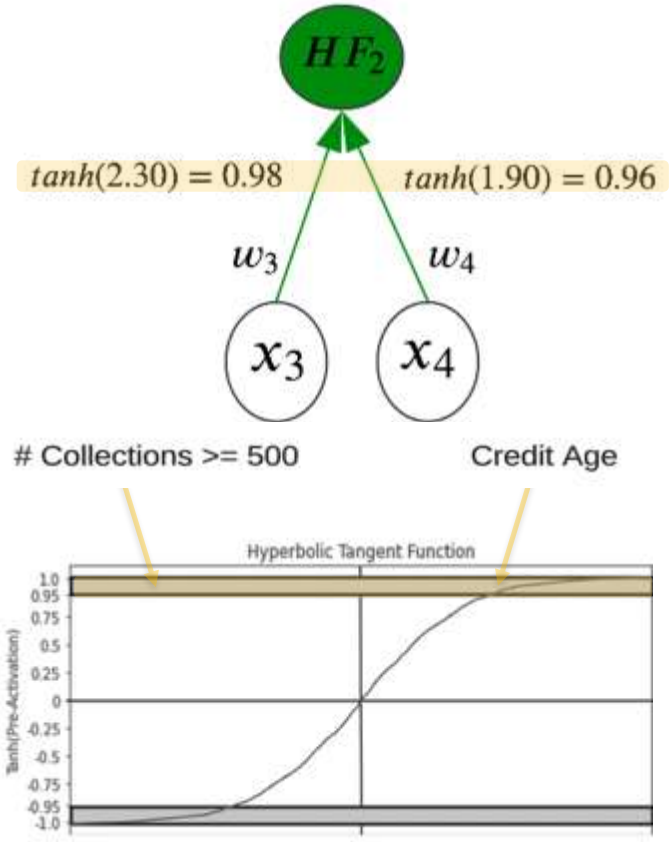


3^6 possible activation states and reasons for the value of HF

- Even for a relatively "simple" HF with 6-input connections, codifying and assigning reasons to each of the 729 activation states would be intractable to any human

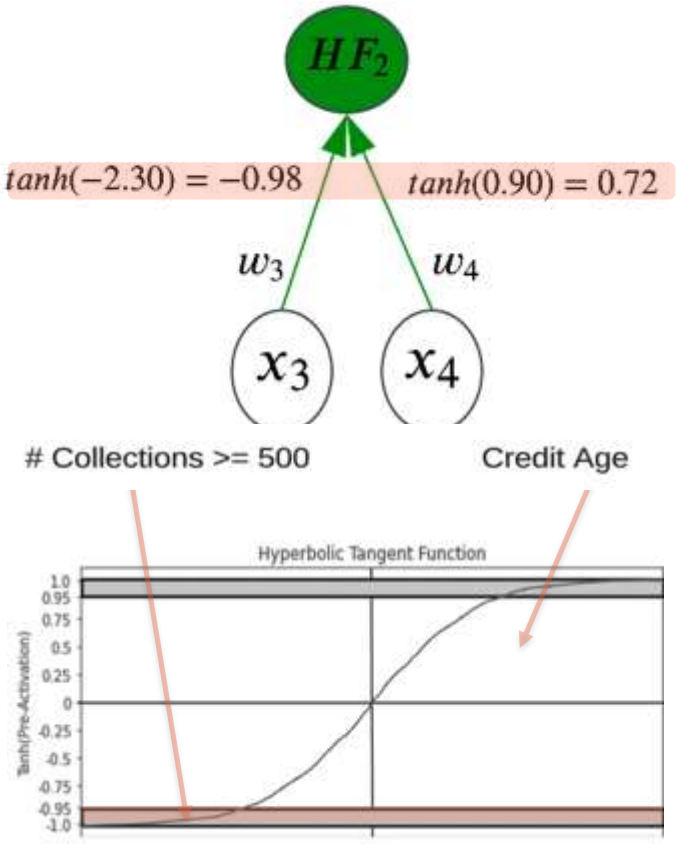
Constructing Reasons for Interpretable Hidden Features

Neural Networks: assigning reasons to activation states, example #1



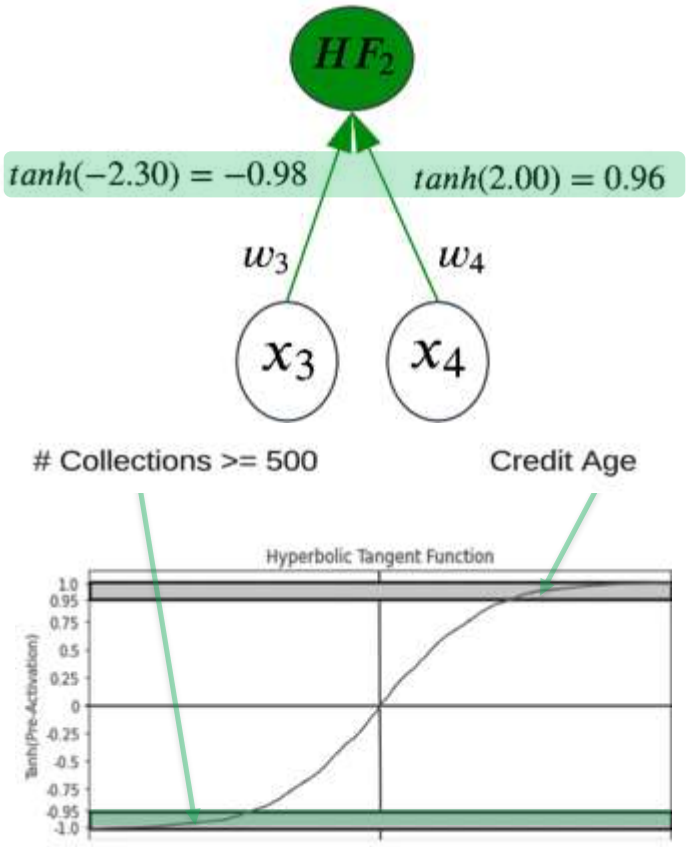
Activation States (x_3, x_4)	Reasons for HF_2
(-1,-1)	"Low # of collections >= $500 and high credit age"
(-1, 0)	"Low # of collections >= $500"
(-1, 1)	If x_3 overwhelms x_4 : "Low # of collections >= $500 offset by low credit age" If x_4 overwhelms x_3 : "Low credit age offset by low # of collections >= $500"
(0, -1)	"High credit age"
(0, 0)	"Typical # of collections >= $500 and typical credit age"
(0, 1)	"Low credit age"
(1, -1)	If x_3 overwhelms x_4 : "High # of collections >= $500 offset by high credit age" If x_4 overwhelms x_3 : "High credit age offset by high # of collections >= $500"
(1, 0)	"High # of collections >= $500"
(1, 1)	"High # of collections >= $500 and low credit age"

Neural Networks: assigning reasons to activation states, example #2



Activation States (x_3, x_4)	Reasons for HF_2
(-1,-1)	"Low # of collections >= $500 and high credit age"
(-1, 0)	"Low # of collections >= $500"
(-1, 1)	If x_3 overwhelms x_4 : "Low # of collections >= $500 offset by low credit age" If x_4 overwhelms x_3 : "Low credit age offset by low # of collections >= $500"
(0, -1)	"High credit age"
(0, 0)	"Typical # of collections >= $500 and typical credit age"
(0, 1)	"Low credit age"
(1, -1)	If x_3 overwhelms x_4 : "High # of collections >= $500 offset by high credit age" If x_4 overwhelms x_3 : "High credit age offset by high # of collections >= $500"
(1, 0)	"High # of collections >= $500"
(1, 1)	"High # of collections >= $500 and low credit age"

Neural Networks: assigning reasons to activation states, example #3 of a mixed state



Activation States (x_3, x_4)	Reasons for HF_2
(-1 ,-1)	"Low # of collections >= $500 and high credit age"
(-1, 0)	"Low # of collections >= $500"
(-1, 1)	If x_3 overwhelms x_4 : "Low # of collections >= $500 offset by low credit age" If x_4 overwhelms x_3 : "Low credit age offset by low # of collections >= $500"
(0, -1)	"High credit age"
(0, 0)	"Typical # of collections >= $500 and typical credit age"
(0, 1)	"Low credit age"
(1, -1)	If x_3 overwhelms x_4 : "High # of collections >= $500 offset by high credit age" If x_4 overwhelms x_3 : "High credit age offset by high # of collections >= $500"
(1, 0)	"High # of collections >= $500"
(1, 1)	"High # of collections >= $500 and low credit age"

Making Neural Networks With Interpretable Hidden Features More Explainable:

- ensuring that a maximum of N-number of latent features can cooperate to provide a response to the incoming input data,
  - N is defined based on industry specification (e.g., 3 to 4 reasons in credit risk)

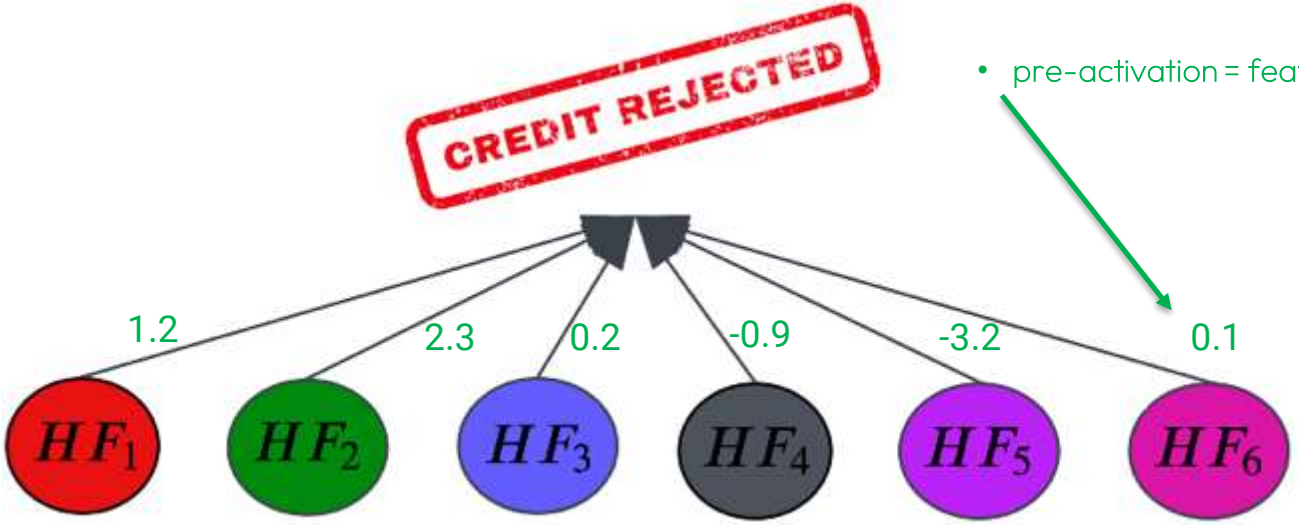
Neural Networks: enforcing additional explainability constraints at the output layer

- In regulatory situations (e.g., credit risk), automated decisions produced by machine learning models are often constrained to no more than 3-4 reasons to provide to consumers to enable less complexity in explanation

- In a densely connected output node within a neural network, all hidden features cooperate to provide a response at the output

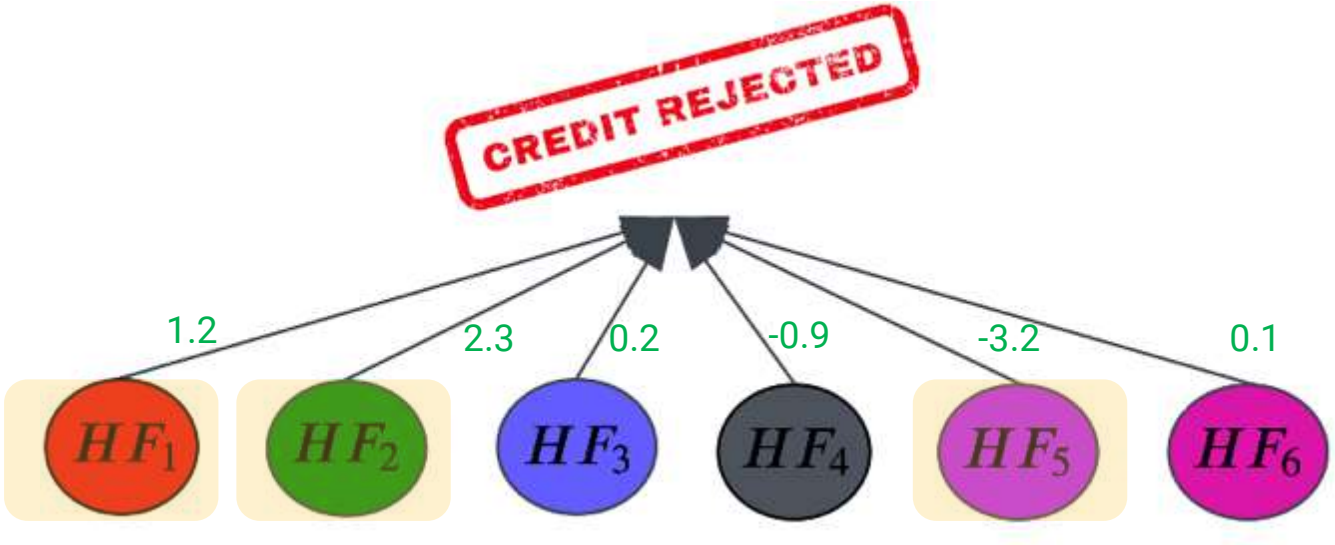
- To bring explainability to the forefront of a neural network, we need to accentuate the learning process of a limited number of hidden features that capture the most significant input interactions when generating a response to the incoming data

Neural Networks: enforcing additional explainability constraints at the output layer



- pre-activation = feature value x weight

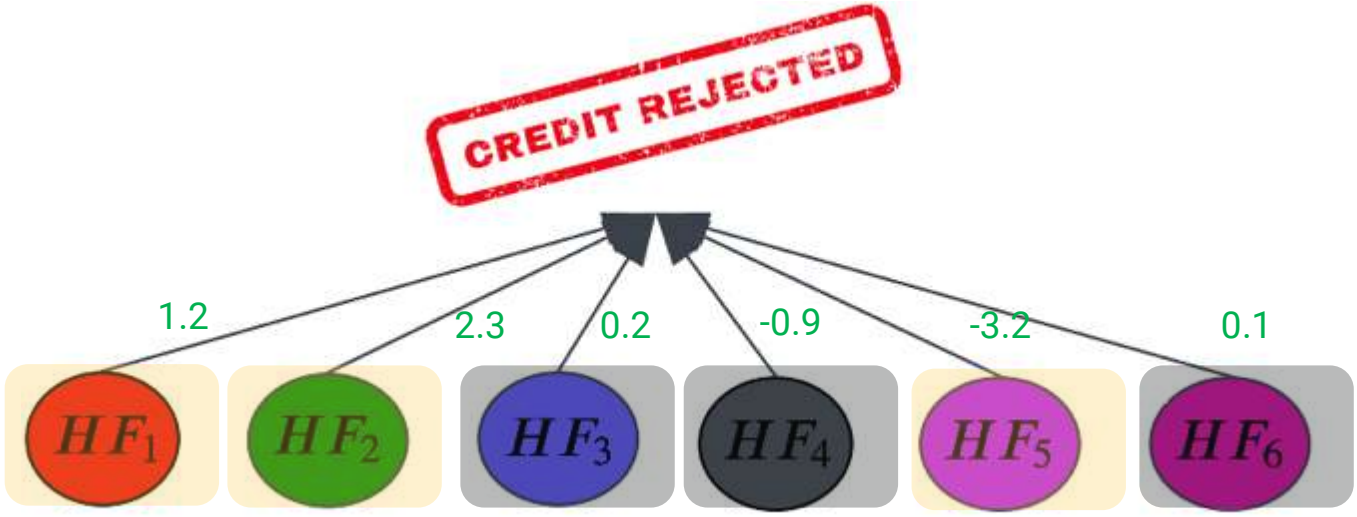
Neural Networks: enforcing additional explainability constraints at the output layer



- Set parameter N to meet the prescribed number of explanations associated with automated decision making based on industry specification of number of reasons. Let's assume N = 3
- Rank order hidden features $HF_5, HF_2, HF_1, HF_4, HF_3, HF_6$ by the magnitude of the pre-activation terms' absolute v

- are the strongest drivers of the output node's score

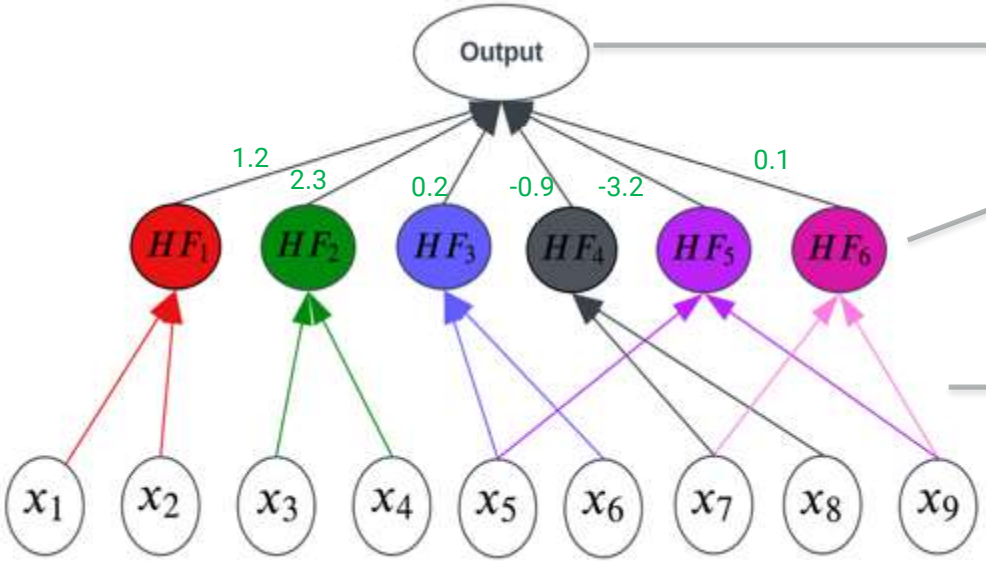
Neural Networks: reinforcing gradients of the reason space dimensionality



pre-activation terms NOT conforming with the explanation space of size N = 3

pre-activation terms conforming with the explanation space of size N = 3

Neural Networks: inference time – generating reasons for top N=3 hidden features



Score = 950 -> high default risk

are the strongest drivers of the output node's score

retrieve activation state matrices

Activation States (x_5, x_6)	Reasons for HF_5
(-1, -1)	xxx
(-1, 0)	xxx
(-1, 1)	xxx
(0, -1)	xx
(0, 0)	xx
(0, 1)	xxx
(1, -1)	xxx
(1, 0)	xxx
(1, 1)	xxx

Activation States (x_3, x_4)	Reasons for HF_2
(-1, -1)	xxx
(-1, 0)	xxx
(-1, 1)	xxx
(0, -1)	xx
(0, 0)	xx
(0, 1)	xxx
(1, -1)	xxx
(1, 0)	xxx
(1, 1)	xxx

Activation States (x_1, x_2)	Reasons for HF_1
(-1, -1)	xxx
(-1, 0)	xxx
(-1, 1)	xxx
(0, -1)	xx
(0, 0)	xx
(0, 1)	xxx
(1, -1)	xxx
(1, 0)	xxx
(1, 1)	xxx

THANK YOU!

krzysztofnalborski@fico.com

1. Neural networks are highly predictive but challenging to interpret and explain.

2. Hidden layer(s) is a key predictive component of a neural network, but fully interpreting and explaining its properties and typically dense connections is very challenging.

3. FICO's innovation addresses the interpretability and explainability issues with a novel neural network training method and construction of human traceable and readable reasons associated with each hidden feature.

Patent Application No. 63/688,584 and Patent Application No. 18/634,652

Appendix:

https://www.fico.com/blogs/deep-dive-how-make-black-box-neural-networks-explainable

https://www.fico.com/blogs/fighting-bias-how-interpretable-latent-features-remove-bias-neural-networks