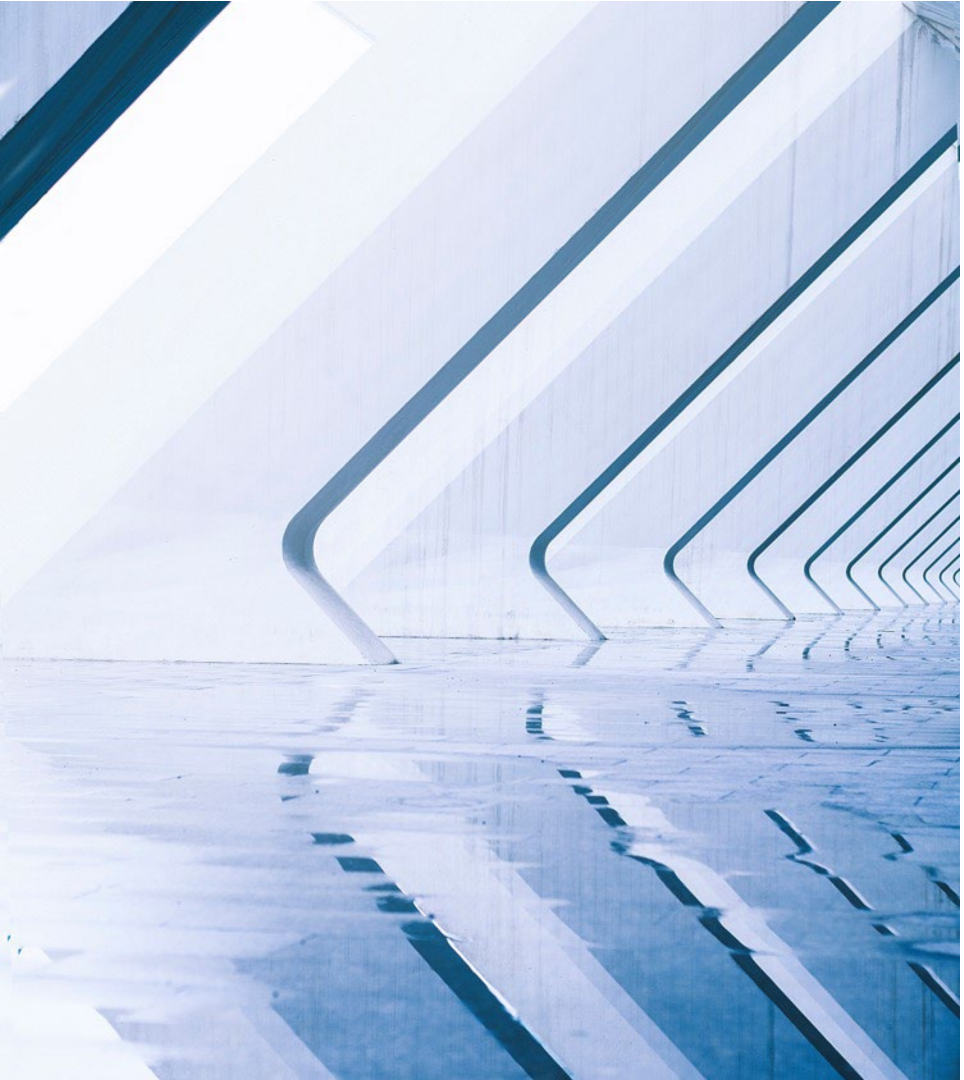


First-to-Saturate Principle for Consistent Explanations of Neural Networks

Krzysztof Nalborski
Lead Scientist
AI Innovation and Development
FICO

1. Neural networks are highly predictive but inherently unexplainable.

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

3. FICO's latest invention tackles the problem of explainability with two key features: a novel first-to-saturate principle and a construction of interpretable hidden nodes.

Neural Networks: highly predictive but very challenging to explain

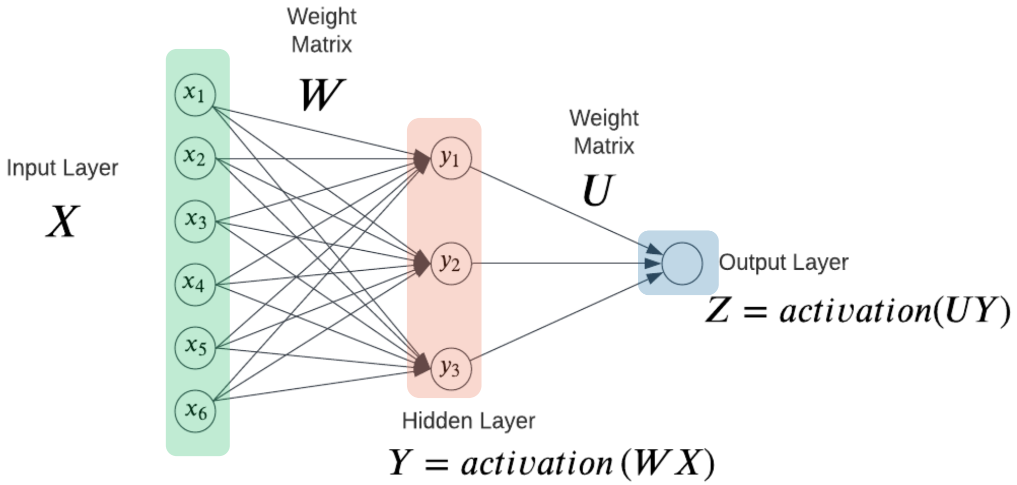
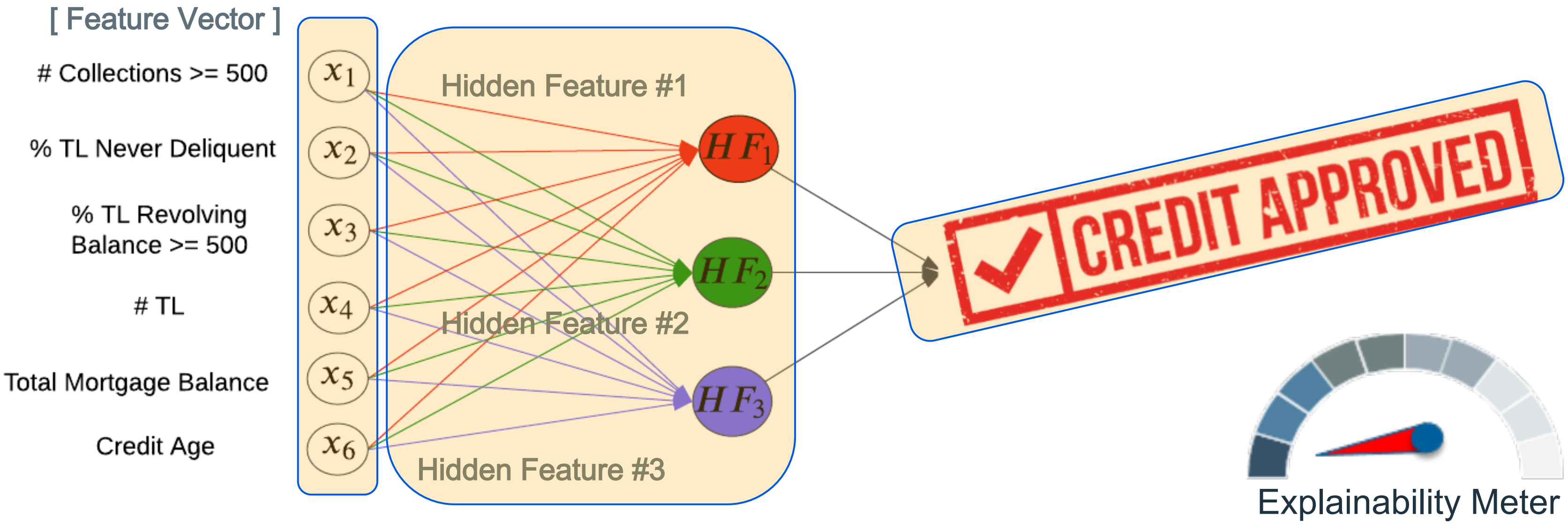


Figure 1. A simple fully connected neural network. Hidden and output layer bias vectors not included for simplification.

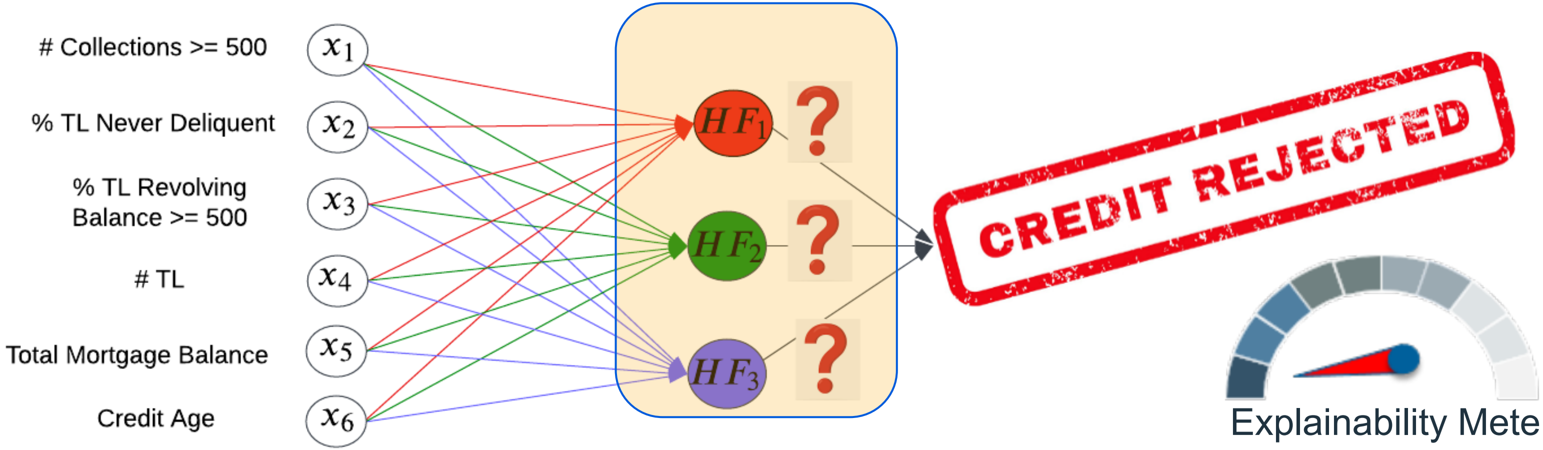
- **Commercially built by FICO for over 30 years** to be used in AI-based decisioning platforms in numerous industries including banking, auto or telco.
- They are typically fully connected containing an input layer, one or more hidden layers, and an output layer.
- **They are highly predictive but very challenging to explain.**
- Hidden layer(s) is the key component of a neural network enabling to model complex and non-linear input data relationships— **understanding and deciphering these relationships have been a focal ExplainableAI (xAI) research area at FICOin the recent years.**

Neural Networks: magical and highly predictive hidden features

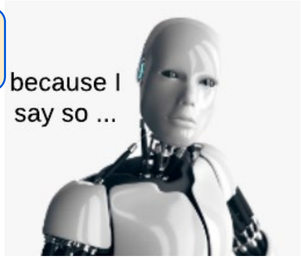


- An **Input Layer** takes domain specific and engineered features and passes them to the rest of the network.
- A **Hidden Layer** (one or more) learns **Hidden (Latent) Features** which are key predictive components enabling a network to discover **complex and non-linear** relationships between the input features.
- An **Output Layer** combines **Hidden (Latent) Features** to produce a score to be used for decisioning.

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.



Explainable AI (xAI): current approaches and challenges

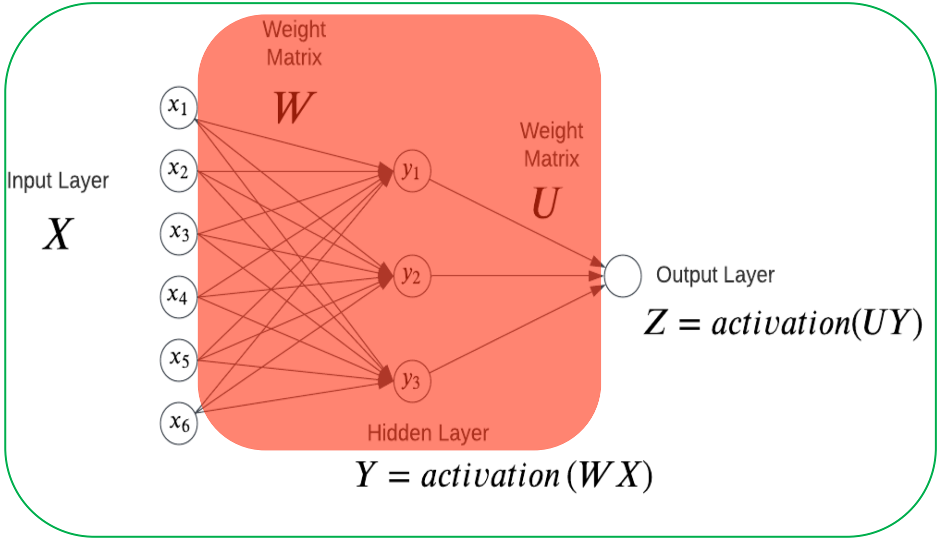


Figure 1. A simple fully connected neural network. Hidden and output layer bias vectors not included for simplification.

Questionable:

- Local Interpretable Model-Agnostic Explanations (LIME)
- Shapley Additive exPlanations (Shapley)

- LIME: Injectnoisy data around point being investigated, score, and train a new linear model. Local decision boundary is investigated.
- Shapley: Calculate Shapley values for all features to make local explanations.
- Both approaches DO NOT consider internals of the original model.

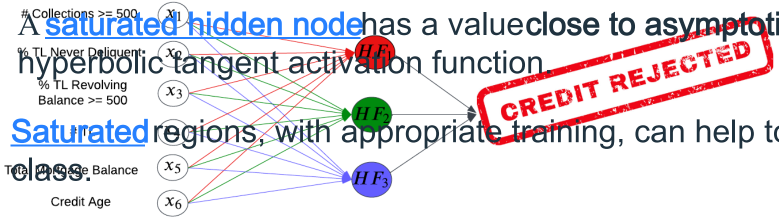
Innovative:

First-to-Saturate (FTS)

- Understanding a neural network's internals to determine which input features drive each (hidden) node into saturation.

First-to-Saturate (FTS) Principle: what is a hidden node saturation?

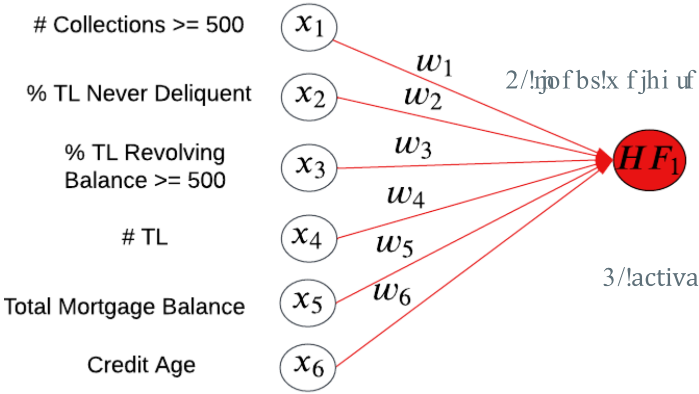
- A saturated hidden node has a value **close to asymptotic ends of an activation function range** For example, close to -1 or 1 for a hyperbolic tangent activation function.



- Saturated regions, with appropriate training, can help to identify the strongest non-linear relationships representative of each class.

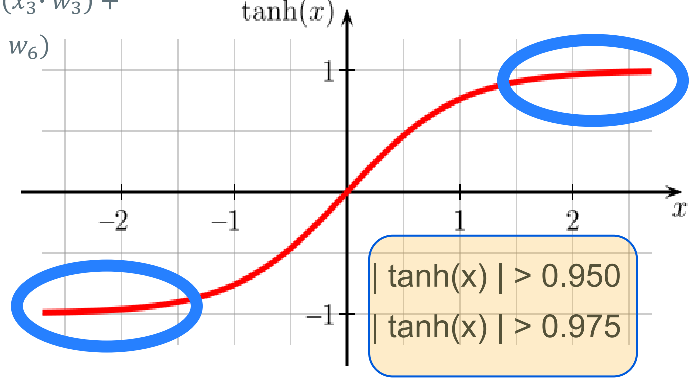
- LWS can keep increasing linearly, but with hyperbolic tangent activation function, its value will be non-linearly transformed to a bounded interval.

  - Which are the most important input-weight combinations and how many of them do we need to push a node into a saturated region? What does saturation mean numerically?



2/!ljofbs!xfjhiufe!tvn!)MXT* $= (x_1 \cdot w_1) + (x_2 \cdot w_2) + (x_3 \cdot w_3) + (x_4 \cdot w_4) + (x_5 \cdot w_5) + (x_6 \cdot w_6)$

3/!activation value $= \tanh(LWS)$

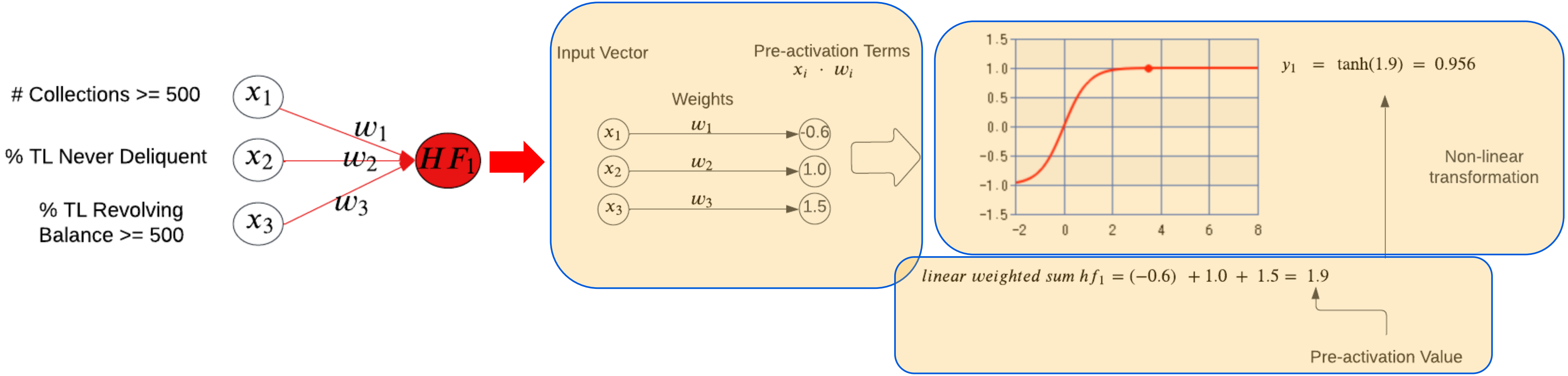


| tanh(x) | > 0.950

| tanh(x) | > 0.975

First-to-Saturate (FTS) Principle: computational paths to saturation > 0.95

- With 3 input features value of **hidden feature HF_1** is 0.956 which is very close to the upper bound of the activation function.



Collections >= 500

% TL Never Deliquent

% TL Revolving Balance >= 500

Input Vector

Weights

Pre-activation Terms $x_i \cdot w_i$

$y_1 = \tanh(1.9) = 0.956$

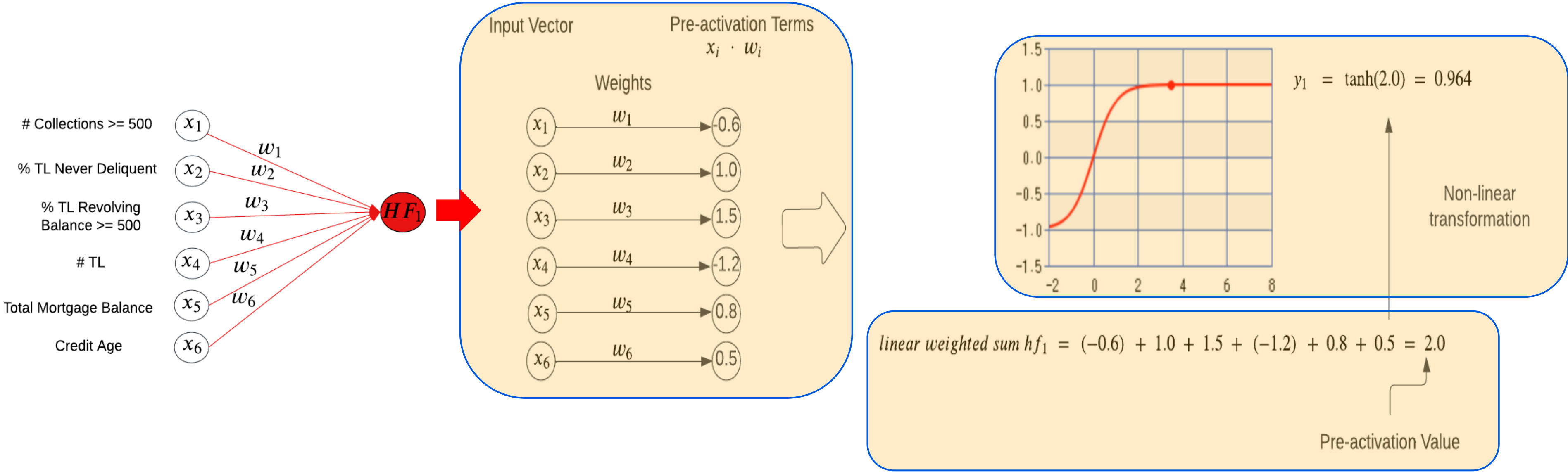
Non-linear transformation

linear weighted sum $hf_1 = (-0.6) + 1.0 + 1.5 = 1.9$

Pre-activation Value

First-to-Saturate (FTS) Principle: computational paths to saturation > 0.95

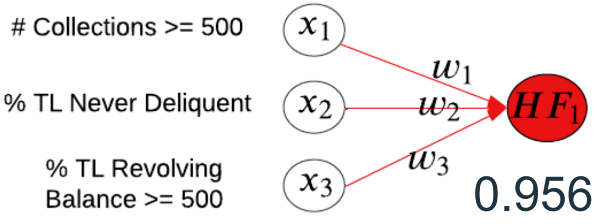
- **With 6 input features, which is 3 more than before,** value of hidden feature HF_1 is 0.964. From reaching the saturation threshold perspective, this value **is already beyond the point regarded as necessary** for the network to learn hidden representation of the strongest input-weight connections incoming into hidden feature HF_1 .



$y_1 = \tanh(2.0) = 0.964$

$\textit{linear weighted sum } hf_1 = (-0.6) + 1.0 + 1.5 + (-1.2) + 0.8 + 0.5 = 2.0$

First - to - Saturate (FTS) Principle: computational paths to saturation > 0.95



Collections >= 500 x_1

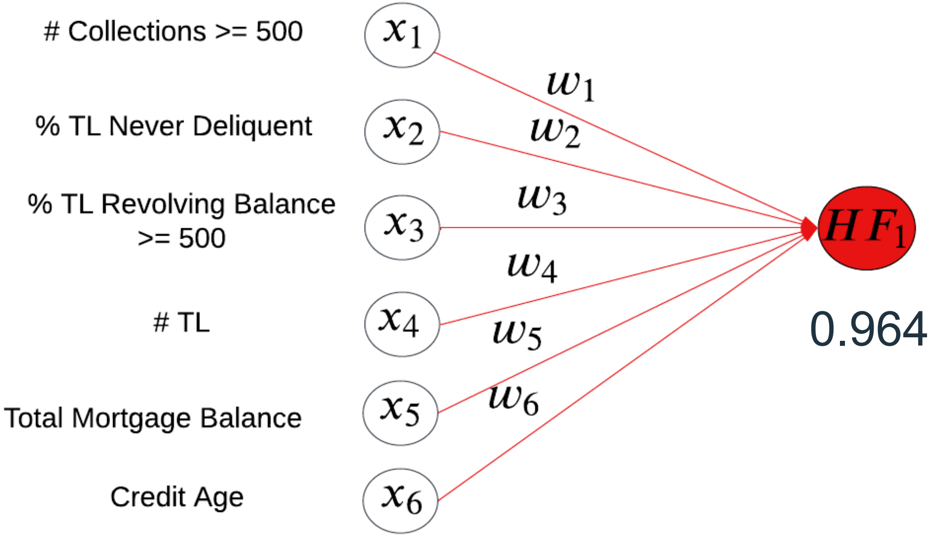
% TL Never Deliquent x_2

% TL Revolving Balance >= 500 x_3

w_1 w_2 w_3 HF_1

0.956

VS



Collections >= 500 x_1

% TL Never Deliquent x_2

% TL Revolving Balance >= 500 x_3

TL x_4

Total Mortgage Balance x_5

Credit Age x_6

w_1 w_2 w_3 w_4 w_5 w_6 HF_1

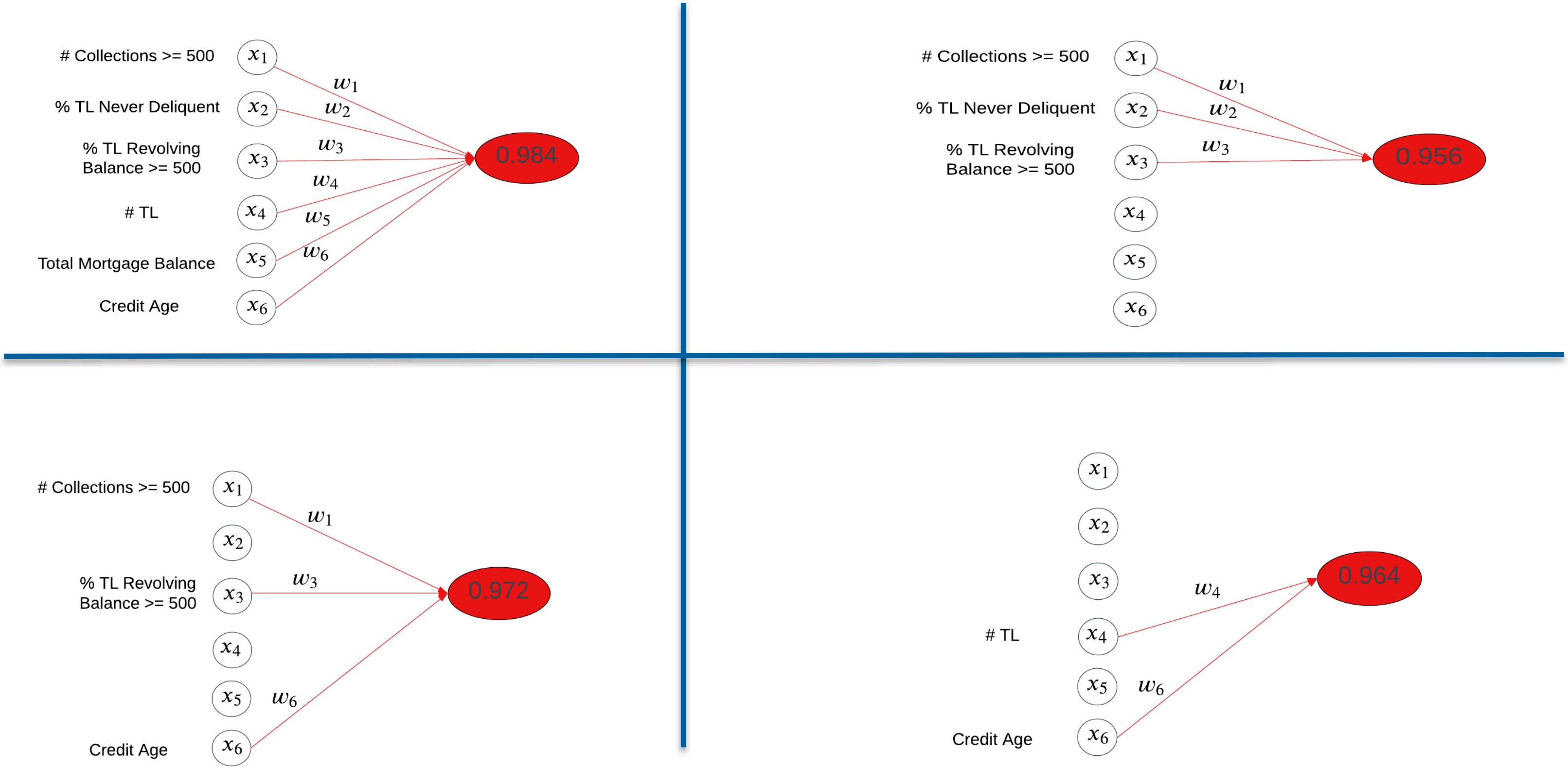
0.964

- Adding 3 extra features increases hidden feature's HF_1 information value, but even without these extra features, value of this hidden feature is **already in the saturated region** and close to the upper bound of the activation function.
- **Saturation (without unnecessary oversaturation) is needed to learn new complex relationships** and reduces the network to a binary state to map inputs to their corresponding labels.
- The new features (x_4, x_5, x_6) only marginally contribute to the node's final value and introduce **additional complexity and unnecessary ambiguity related to understanding and explaining the** hidden feature HF_1 that the neural network learns through training.

FICO

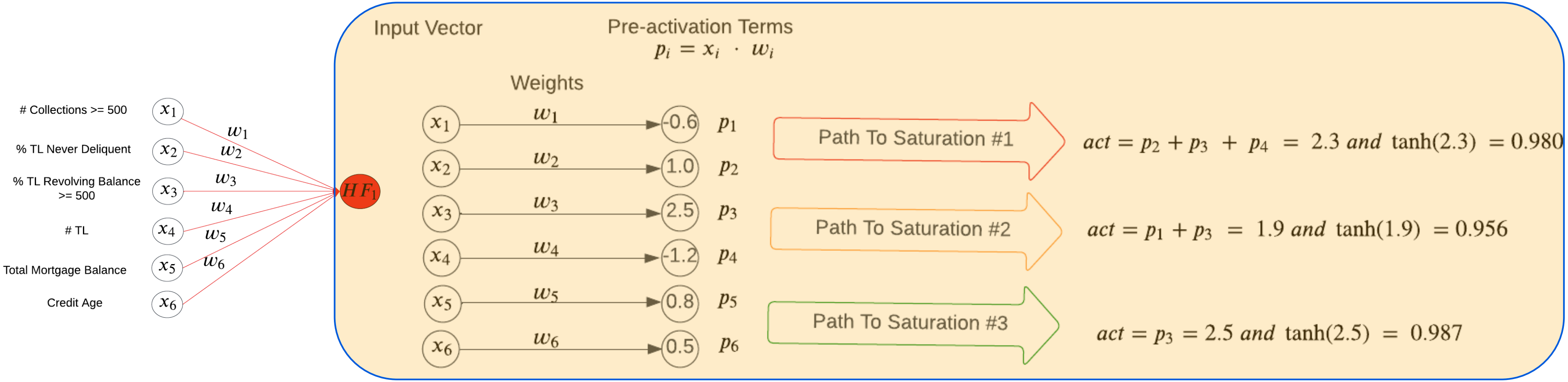
First-to-Saturate (FTS) Principle: saturation modes

- Hidden node HF$_1$ can have numerous and non-deterministic computational paths that lead to **different saturation modes** during training and often **only a subset of features is needed to reach saturation. FTS algorithm can deterministically find that subset** and **rank order** the features.



First-to-Saturate (FTS) Principle: finding ranked features that push a node into saturation

- **FTS algorithm**algorithm allows to define a saturation threshold. For example, a hidden node's absolute activation value > 0.95:
  - $|y_i| > 0.95$



$act = p_2 + p_3 + p_4 = 2.3$ *and* $\tanh(2.3) = 0.980$

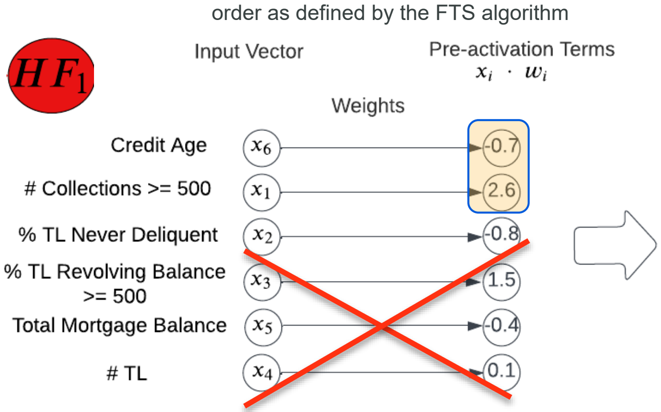
$act = p_1 + p_3 = 1.9$ *and* $\tanh(1.9) = 0.956$

$act = p_3 = 2.5$ *and* $\tanh(2.5) = 0.987$

- Based on magnitude-sorted by their absolute value pre-activation terms, for the entire training data corpus and for each hidden node, **FTS algorithm finds most statistically likely lists of ranked features that can push a node into saturation**

- For example, pre-activation terms of **features {x_6, x_1, x_2} in this order** are most likely to push **hidden feature HF_1** into saturation because they led to 95%+ of all saturations during training.

First - to - Saturate (FTS) Principle: inference



order as defined by the FTS algorithm

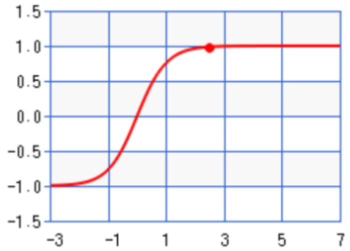
HF_1

Input Vector

Pre-activation Terms
$x_i \cdot w_i$

Weights

Credit Age	x_6	-0.7
# Collections >= 500	x_1	2.6
% TL Never Deliquent	x_2	-0.8
% TL Revolving Balance >= 500	x_3	1.5
Total Mortgage Balance	x_5	-0.4
# TL	x_4	0.1



$$linear\ weighted\ sum\ hf_1 = (-0.7) + 2.6 = 1.9$$

Subset or ranked inputs up until a sum of their pre-activation terms pushes a node over a saturation threshold

$$hf_1 = \tanh(1.9) = 0.956$$

Saturation threshold reached with 2 out of 6 input features

- These 2 features are:
  - Credit Age
  - # Collections >= 500

- Sum of pre-activation terms **associated with features x_6 and x_1 already reaches our saturation threshold**
- The remaining pre-activation terms, for example the **negative -0.8 contribution coming from x_2 , is not included in calculation of the activation value** according to the FTS principle. Some input-weight connections may even be completely masked to simplify a network’s structure if they are found to never lead to saturation.

FICO

First - to - Saturate (FTS) Principle: network training and creation of interpretable hidden nodes

Densely connected network with 144 unique features and 15 hidden nodes

[(0, [0,1,2,3,4,5,…, 143]),
(1, [0,1,2,3,4,5,…, 143]),
(2, [0,1,2,3,4,5, …, 143]),
(3, [0,1,2,3,4,5, …, 143]),
(4, [0,1,2,3,4,5, …, 143]),
(5, [0,1,2,3,4,5, …, 143]),
(6, [0,1,2,3,4,5, …, 143]),
(7, [0,1,2,3,4,5, …, 143]),
(8, [0,1,2,3,4,5, …, 143]),
(9, [0,1,2,3,4,5, …, 143]),
(10, [0,1,2,3,4,5, …, 143]),
(11, [0,1,2,3,4,5, …, 143]),
(12, [0,1,2,3,4,5, …, 143]),
(13, [0,1,2,3,4,5, …, 143]),
(14, [0,1,2,3,4,5, …, 143])]

- hidden node at index 0 and all remaining hidden nodes have incoming connections from all 144 input features; all 144 features are used to compute each hidden feature's value

- During network training with the FTS principle, only weights corresponding to the most "active" input features are updated.
- FTS aims to find a subset of ranked input features minimally sufficient for a hidden node to saturate.
- Training with FTS simplifies a network's structure- hidden layer's weight matrix is masked to only allow feature combinations already proven to be relevant based on their "driving" of hidden nodes into saturation.

VS

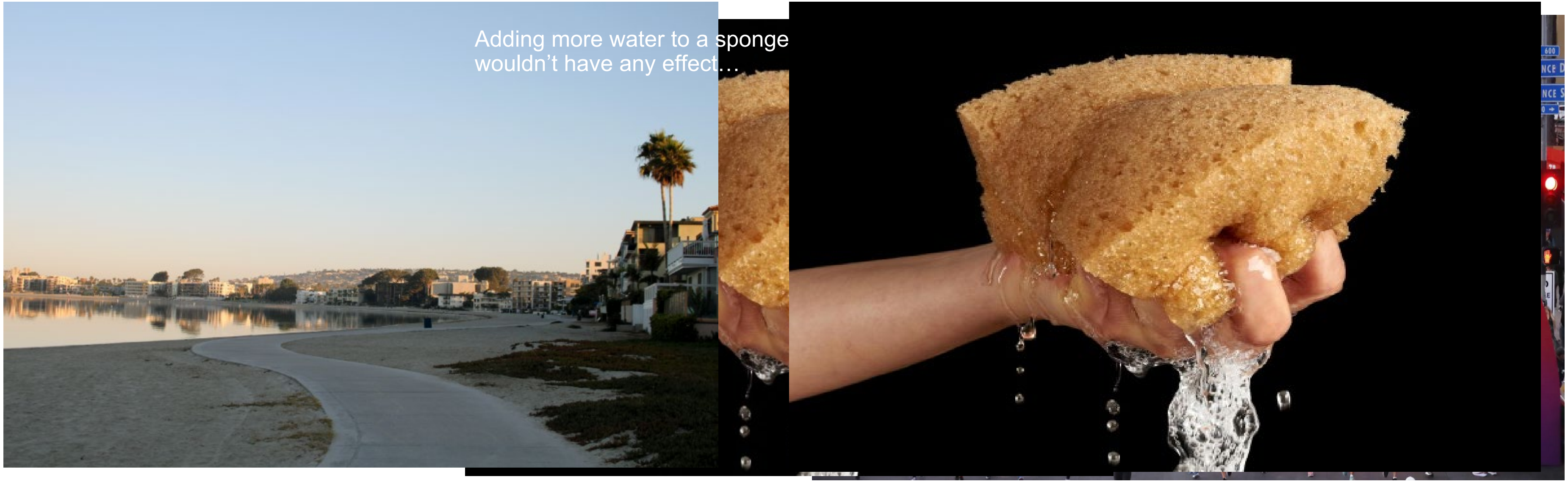
Sparsely connected network trained with the FTS principle

[(0, [14, 69, 71, 34, 48, 37, 26]),
(1, [37, 100, 81, 22, 88, 120, 34]),
(2, [119, 14, 56, 23, 56, 77, 23]),
(3, [113, 44, 136, 23, 48, 66, 44]),
(4, [116, 38, 31, 119, 35, 91, 111]),
(5, [117, 76, 31, 14, 67, 78, 65]),
(6, [121, 56, 7, 138, 12, 24, 77]),
(7, [41, 100, 83, 99, 144, 108]),
(8, [114, 129, 107, 52, 100, 104),
(9, [140, 22, 4, 8, 42, 127, 88, 91]),
(10, [69, 14, 56, 62, 120, 105]),
(11, [119, 44, 31, 37, 76, 44, 12]),
(12, [136, 4, 21, 117, 80, 130, 1]),
(13, [83, 14, 0, 11, 53, 108, 22]),
(14, [123, 100, 32, 89, 98, 90])]

- hidden node at index 0, only connects with features at certain indices; same for the remaining hidden nodes

First-to-Saturate (FTS) Principle: a different take on saturation

- **saturation** simply means filling a thing or a place with “something” to an extent where there is **enough** of that ”something” and more of that “something” would **have no additional effect** on that thing or that place …



Adding more water to a sponge wouldn’t have any effect…

- **What saturates Krzysztof’s level of endorphins / what makes Krzysztof happy while running?**
  - Location (by the beach, around a lake, in a canyon).
  - Time of the day (early morning).
  - Style of running (progression run, intervals).

3 factors (and their combinations) to put Krzysztof in one of the happiness related **saturation MODES** while running.

1. Neural networks are highly predictive but inherently unexplainable.

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

3. FICO’s latest invention tackles the problem of explainability with two key features: a novel first-to-saturate principle and a construction of interpretable hidden nodes.

THANK YOU!

krzysztofnalborski@fico.com