www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(3), 2026, 37-52. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i3.1331

Mitigating the Saturation Gap: Inverse Prime-Density Scaling for High-Stakes Probabilistic Modeling

Jaime Aguilar-Ortiz[1,2], Manuel Francisco Gutiérrez Salina[1], Carlos R. Domínguez Mayorga[2]

[1] Universidad Politécnica Metropolitana de Hidalgo (UPMH), Mexico.
[2] Universidad Politécnica de Pachuca, Mexico.
E-mails: jao@upp.edu.mx, derivado29@gmail.com

Abstract. Standard output gates under severe class imbalances expose a limitation of standard probabilistic gates that leave the problem to threshold optimization for rare classes. In regulated environments where frequent threshold adjustments raise governance and authority concerns, this approach becomes problematic. We introduce an output gating function that preserves gradient sensitivity in high-activation regimes during the training period. This study proposes a logarithmically moderated function based on the Prime Number Theorem (PNT), which preserves gradient sensitivity, mitigating the saturation gap that prevents a model from ignoring rare events. Results are demonstrated through a comparison of different output gates and 63 architectural combinations where PrimeSigmoid consistently dominates across hidden representations showing significant gains in Recall and Matthews Correlation Coefficient without degrading AUC or LogLoss calibration. This demonstrates a robust classification for high-stakes financial risk modeling.
Keywords: Sigmoid, Prime Number Theorem, Financial Risk Modeling, Recall, Logarithmic Moderation.

Article Info
Received Ene 26, 2026
Accepted Mar 11, 2026

1 Introduction

Machine learning systems are frequently deployed in settings dominated by extreme class imbalance, including financial risk modeling, fraud detection, and other rare-event prediction tasks. Under these conditions, the probabilistic mapping applied at the output layer becomes a central component of the decision process. The most well-known and widely used output gate remains the logistic sigmoid, defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{1}$$

This well-established and adopted output gate exhibits a rapid saturation for moderately large values of |x|, leading to vanishing gradients and conservative decision behavior (Glorot and Bengio 2010; Rumelhart et al., 1986). This saturation tends to adjust predictions toward the majority class, which results in less false positives but more critical events missed.

The theoretical proposal and experiments of this work are based on key assumptions. First, the overall accuracy is usually misleading since metrics such as F1-score, recall and Matthews Correlation Coefficient are more informative (Chicco & Jurman, 2020). Second, a missing risk event classified as a false negative is more costly than a false alarm (Elkan, 2001).

Motivated by these constraints, the output nonlinearity is modified to reflect a logarithmic moderation inspired by the asymptotic structure governing the density of prime numbers (Hardy & Wright, 1979). This construction replaces exponential saturation with a logarithmically decaying sensitivity, yielding a probabilistic gate whose response remains informative for large-magnitude pre-activations. The resulting behavior offers a structurally grounded alternative aligned with the operational demands imposed by extreme class imbalance.

2 Related Work

This study situates the classification problem within the interaction between probabilistic output mappings and internal neural representations. Rather than treating these components independently, the analysis considers how the choice of output nonlinearity governs probability formation under severe class imbalance, while hidden representations modulate the geometry of the decision boundary. Within this framework, baseline mechanisms are defined explicitly to contextualize the proposed contribution.

A range of probabilistic output gates has been employed historically to map a real-valued logit z into the unit interval under a Bernoulli likelihood. Under the binary cross-entropy formulation, the logistic sigmoid remains the canonical choice, defined as

$$\sigma(x) = \frac{1}{1 + e^{-z}}. \tag{2}$$

Despite its ubiquity, this transformation saturates exponentially for moderate values of $[z]$, leading to vanishing gradients and conservative decision behavior in imbalanced regimes. As a result, models often minimize loss by favoring the majority class, achieving high accuracy while exhibiting poor recall for rare events (Rumelhart et al., 1986).

Several variants have been proposed to mitigate this behavior without altering the loss formulation. Temperature-scaled sigmoids introduce a tunable parameter T,

$$\sigma_T(z) = \frac{1}{1 + e^{-\frac{z}{T}}}. \tag{3}$$

allowing saturation to be delayed or accelerated depending on the temperature (Guo et al., 2017). While such scaling modifies curvature locally, it preserves the exponential decay governing the tails, resulting in only incremental changes in minority sensitivity. Similarly, symmetric alternatives such as the rescaled hyperbolic tangent,

$$f(z) = \frac{1 + \tanh(z)}{2}. \tag{4}$$

share the same saturation mechanism, differing primarily in gradient behavior near the origin due to zero-centered dynamics prior to rescaling (LeCun et al., 2012).

Probabilistic gates derived from alternative noise assumptions have also been explored. The Probit function, based on the cumulative distribution function of the standard normal distribution,

$$\Phi(z) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right]. \tag{5}$$

has long-standing use in econometrics (Bliss, 1934). Although smoother around zero, its tail behavior closely mirrors that of the sigmoid, limiting its effectiveness under extreme imbalance. At the opposite end of the design spectrum, HardSigmoid provides a piecewise linear approximation,

$$f(z) = \text{clip}(0.2z + 0.5,0,1). \tag{6}$$

trading smoothness for computational efficiency (Courbariaux et al., 2015). The presence of extended flat regions with zero gradient, however, leads to failure modes in imbalanced settings, where models collapse into majority-class predictions.

To disentangle the contribution of the output gate from internal feature transformations, performance is evaluated across a diverse set of hidden activation functions that reflect the evolution of nonlinear design in neural networks. Classical choices such as ReLU,

$$\phi(z) = \max(0, z). \tag{7}$$

and its leaky variant,

$$\phi(z) = \max(\alpha z, z)\,. \tag{8}$$

introduce piecewise linear behavior with controlled gradient flow (Nair & Hinton, 2010; Maas et al., 2013). Smooth alternatives including ELU,

$$\phi(z) = z \text{ if } z > 0, \text{else } \alpha(e^z - 1) \text{ if } z \leq 0\ . \tag{9}$$

and Tanh,

$$\phi(z) = \tanh(z)\,. \tag{10}$$

provide saturation in the negative regime while maintaining differentiability (Clevert et al., 2016; LeCun et al., 2012). More recent designs such as GELU,

$$\phi(z) = z \cdot \Phi(z)\,. \tag{11}$$

Swish,

$$\phi(z) = z \cdot \sigma(z)\,. \tag{12}$$

and Mish,

$$\phi(z) = z \cdot \tanh(\ln(1 + e^z))\,. \tag{13}$$

introduce smooth, non-monotonic behavior that improves gradient propagation in deep networks (Hendrycks & Gimpel, 2016; Ramachandran et al., 2017; Misra, 2019). By fixing these internal representations, any systematic performance variation observed can be attributed to the output nonlinearity rather than to hidden-layer expressiveness.

3 Theoretical Foundations of Primesigmoid and PrimeTanhGate

The proposed PrimeSigmoid and PrimeTanhGate are grounded in concepts from analytic number theory. Let $\pi(x)$ denote the prime-counting function,

$$\pi(x) \coloneqq \#\{p \leq x : p \text{ is prime}\}\,. \tag{14}$$

which counts the number of primes less than or equal to x . The Prime Number Theorem establishes the asymptotic relation

$$\pi(x) \sim \frac{x}{\ln x}\,. \tag{1}$$

Heuristic derivations of this density arise from assuming approximate independence of divisibility events, leading to

$$P(n \text{ is prime}) \approx \prod_{p \leq \sqrt{x}} \left(1 - \frac{1}{p}\right). \tag{16}$$

Taking logarithms and applying the first-order Taylor expansion $\ln(1\text{-}u) \approx -u$::

$$\ln\left(\prod_{p\le z}\left(1-\frac{1}{p}\right)\right)\approx-\sum_{p\le z}\frac{1}{p}. \tag{17}$$

By Mertens' theorem,

$$\frac{\sum_{\mathrm{P}\le \mathrm{z}}1}{\mathrm{p}}\approx \mathrm{lnlnz}+\gamma,\ \text{where}\ \gamma\approx 0.577\,. \tag{2}$$

where $\gamma \approx 0.577$is the Euler–Mascheroni constant, yielding a local density proportional to $1/\ln\ x$when $z=\sqrt{x}$. Deviations from this approximation are known to arise from oscillatory terms associated with the non-trivial zeros of the Riemann zeta function (Riemann, 1859).

Justification From Prime Asymptotics to Controlled Saturation: A Probabilistic Response, Gradient Decay Bound and Asymptotic Saturation Rate

The typical logistic sigmoid exhibits exponential saturation, which in severe imbalanced classes creates a problem of dominance that drives the logits toward a huge region that leads the gradient to collapse to zero (Glorot and Bengio 2010, Lin et al, 2017). This behavior neglects minority classes receiving fewer corrective updates during the training period, leading to a low-recall-high accuracy.

The resulting regime is not linear nor abrupt, but sublinear that also grows towards infinity, while the slope that approaches to zero. Applying L'Hôpital's rule,

$$\lim_{n\to\infty}\frac{x}{\ln x}=\lim_{n\to\infty}\frac{\frac{d}{dx}x}{\frac{d}{dx}\ln x}=\lim_{n\to\infty}\frac{1}{1/x}=\lim_{n\to\infty}x=\infty\,. \tag{18}$$

This principle is directly relevant since it's the key design to modulate the typical disadvantage of the logistic sigmoid under a high-stakes classification task. The proposal is therefore not a mere alteration of the activation function, but it directly counteracts the structural tendency to ignore a minority class.

By replacing the linear growth in the exponent, the new argument grows proportionally to $z/log\lfloor z\rfloor$. This modification preserves probabilistic interpretability and monotonicity while for the saturation means that instead of having an exponential decay, the gradient will vanish logarithmically, and therefore still sustaining a learning curve for a minority class.

Encapsulating this idea in a sigmoidal map, would still be confined in a unit interval $[0,1]$, and it also decouples probabilistic calibration from saturation dynamics. This separation enhances the sensitivity to minority classes without sacrificing AUC or LogLoss.

Figure 1 compares the probabilistic response of the basic binary output gates considered in this document. Probit functions and standard sigmoids saturate rapidly, while temperature scaling just shifts the transition without changing the exponential feature. Conversely the PrimeSigmoid exhibits a moderate growth zone, which promotes a regime that reconsider the discrimination boundary by preserving gradient sensitivity. The preactivation graph dominium for $z\in[-10{,}20]$, $\Delta z\approx .05$ is shown below:

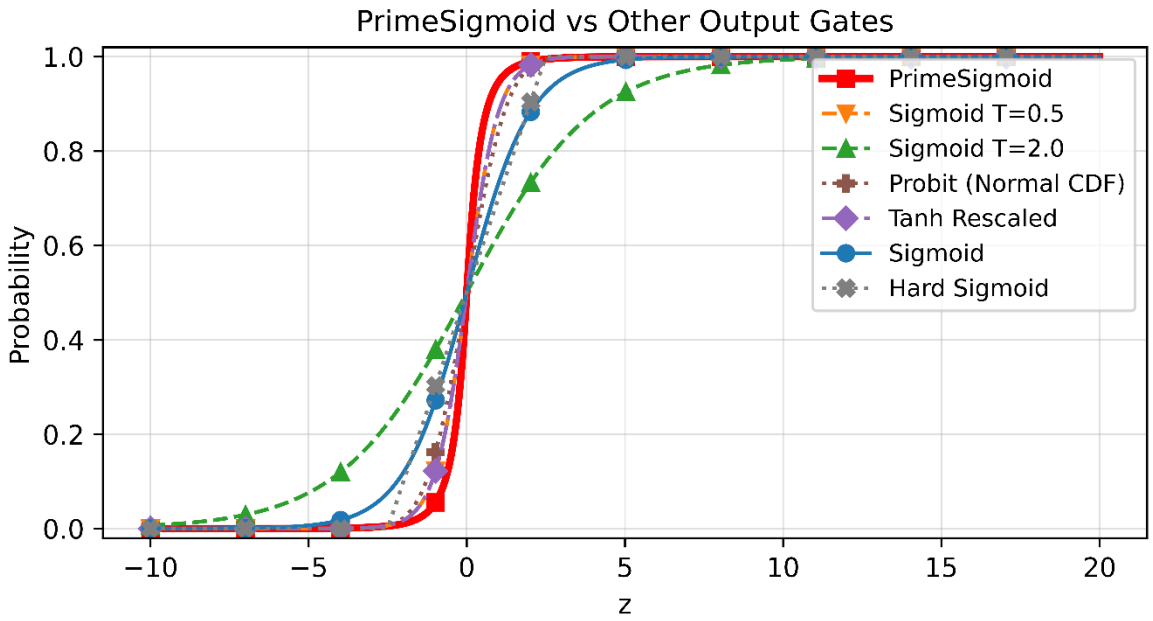

Fig. 1. Output Gates Probabilistic Behaviour

Implying that the asymptotic density of prime numbers decays as $\frac{1}{\ln x}$ ((Hardy and Wright, 1979). The PrimeSigmoid directly incorporates this inverse logarithmic scaling into its argument. Specifically, defining the scaling function g(x) as:

$$g(x) = \frac{\pi x}{\ln(2 + |x| + \varepsilon)}. \tag{19}$$

The multiplicative factor π is introduced as a neutral scale value that preserves moderation while maintaining sensitivity. It is proposed only to attenuate the effect of logarithmic normalization to ensure that the gate remains comparable in steepness to the sigmoidal map. Therefore, π is merely a calibration constant adjusting the scale without any other modification to the structure.

Additive number 2 inside the term serves as a bound for positive values and guarantees positivity of denominator as $x \in \mathbb{R}$ ensuring continuity and differentiability of the scaling function across the domain. The choice of 2 gives us not only a well-defined and bounded normalization factor but also prevents excessive amplification of small preactivations.

Also observe that for large |x|:

$$g(x) \sim \frac{\pi x}{\ln|x|}. \tag{20}$$

which grows sub linearly with respect to x. This contrasts sharply with the linear growth of the standard sigmoid argument (x) and forms the mathematical core of the proposed activation.

The resulting construction yields the PrimeSigmoid activation f(x), formulated as a logarithmically moderated transformation of the classical logistic sigmoid. By embedding inverse logarithmic scaling into the argument governing saturation, the function preserves the probabilistic interpretation of the standard sigmoid while altering its asymptotic response to large-magnitude inputs.

$$f(x) = \sigma\big(g(x)\big) = \frac{1}{1 + \exp\left(-\frac{\pi x}{\ln(2 + |x| + \varepsilon)}\right)}. \tag{21}$$

where $\sigma(\cdot)$ denotes the standard logistic sigmoid, $\pi \approx 3.14159$, an, and $\varepsilon = 10^{-8}$ is introduced for numerical stability. The additive constant inside the logarithm $(2 + |x|)$) ensures the denominator is strictly positive for all real x. Specifically, at $x = 0$, the denominator evaluates to $n(2 + \varepsilon) \approx 0.693$, ensuring a smooth transition around the origin.

The derivative of the PrimeSigmoid follows from the chain rule:

$$f'(x) = \sigma'\big(g(x)\big) \cdot g'(x)\,. \tag{22}$$

Recalling that $\sigma'(u) = \sigma(u)\big(1\text{-}\sigma(u)\big) \leq 0.25$, it f , it follows that $f'(x) \leq 0.25g'(x)$. The derivative of the inner function $g(x)$ satisfies, for large $|x|$:

$$g'(x) \sim \frac{\pi}{(\ln\ |x|)^2}\,. \tag{23}$$

This implies that the gradient approaches zero only logarithmically. This slow decay stands in distinct contrast to the exponential decay of the standard sigmoid ($O\big(e^{-|x|}\big)$)) (Rumelhart et al., 1986), whose gradient rapidly collapses for large-magnitude inputs. Consequently, PrimeSigmoid preserves non-negligible gradient sensitivity even in regions of high pre-activation magnitude.

As $x \to +\infty$:

$$f(x) \to 1, \quad 1 - f(x) \sim \exp\left(-\frac{\pi x}{\ln x}\right). \tag{24}$$

This indicates a significantly heavier tail than that of the conventional sigmoid. This behavior ensures that the function approaches saturation limits more gradually, maintaining probabilistic resolution for extreme inputs.

We can extend the same Primesigmoid principle to the hidden layers via the PrimeTanhGate, adapting the architectural philosophy of Mish.

This section demonstrates that the connection is not heuristic, its mathematical consequences are reflected in improved gradient behavior and enhanced minority recall, under imbalanced classes.

Prime-Gated Hidden Units

While the PrimeSigmoid serves as the probabilistic output, the internal representation requires a mechanism to manage information flow. Drawing architectural inspiration specifically from the self-regularized non-monotonicity of the Mish activation (Misra, 2019), we introduce the PrimeTanhGate.

Unlike Mish, which relies on a softplus-tanh core, our proposal replaces the internal scaling with the logarithmic moderation.

PrimeTanhGate

Let the pre-activation of a hidden unit be denoted by $z \in R$. We define the PrimeTanhGate by first introducing a logarithmically moderated prime core transformation, $c(z)$, inspired by the asymptotic density of prime numbers:

$$c(z) = \frac{\pi z}{\ln(2 + |z| + \varepsilon)}\,. \tag{25}$$

where $\varepsilon > 0$ is a small constant (10^{-8}) introduced for numerical stability. The inclusion of the constant 2 inside the logarithm guarantees that the denominator remains strictly positive for all real inputs, avoiding singular behavior near the origin.

This prime core transformation preserves the sign of the input z and introduces a sublinear growth regime for large magnitudes. Building on this core, the PrimeTanhGate activation is defined as a multiplicative gated function:

$$\phi_{\text{PrimeTanh}}(z) = z \cdot \mathcal{G}(z)\,. \tag{26}$$

where the gating function $\mathrm{G(z)}$ is defined by:

$$\mathcal{G}(z) = \frac{1 + \tanh\big(c(z)\big)}{2}. \tag{27}$$

The gating function $\mathrm{G(z)} \in (0,1)$ introduces a smooth and bounded modulation of the input signal while preserving the sign of the activation across its entire domain. By construction, the transformation avoids hard saturation and maintains continuity, ensuring that the resulting activation remains differentiable and well-behaved during optimization. In regimes where the gate remains fully open, the explicit linear dependence on $z$allows the activation to reduce to an identity mapping, preventing distortion of informative signals.

This identity-preserving behavior is complemented by a smoothly varying gating component derived from the hyperbolic tangent, which regulates transitions without introducing discontinuities or flat regions associated with hard thresholding. Stability of gradients is therefore maintained even as the gate progressively constrains the signal. Superimposed on this structure, logarithmic moderation inherited from the prime-inspired core $c(z)$ governs growth through inverse logarithmic scaling, providing a principled mechanism to delay saturation while retaining sensitivity for large-magnitude inputs.

Derivative Analysis

Near the Origin: Using Taylor expansions where $\ln(2 + |\mathrm{z}|) \approx \ln 2$, the prime core behaves approximately linearly as $\mathrm{c(z)} \approx \mathrm{kz}$ with $\mathrm{k} = \frac{\pi}{\ln 2}$. The activation can be locally approximated as:

$$\phi_{\text{PrimeTanh}}(z) \approx \frac{z}{2} + \frac{k}{2} z^2\,. \tag{28}$$

This reveals a finite slope of approximately 0.5 at the origin, ensuring stable gradients during initialization.
Gradient Stability: From a differential perspective, the derivative follows the product rule $\phi'(\mathrm{z}) = \mathrm{G(z)} + \mathrm{z} \cdot \mathrm{G}'(\mathrm{z})$. Since $\mathrm{G(z)} \in (0,1)$ and its derivative is bounded, the resulting gradient is continuous and free from sharp peaks, aligning with the empirical stability required for training deep architectures.

Asymptotic Behavior

The asymptotic behavior of PrimeTanhGate is instructive. As $\mathrm{z} \to +\infty$, the logarithmic term satisfies $\ln(2 + |\mathrm{z}|) \sim \ln \mathrm{z}$. . Consequently, $\mathrm{c(z)} \to +\infty$, , causing $\tanh\big(\mathrm{c(z)}\big) \to 1$ and $(\mathrm{z}) \to 1$. In this regime, the activation converges to:

$$\phi_{\text{PrimeTanh}}(z) \approx z\,. \tag{29}$$

This recovers the identity mapping with a non-vanishing gradient, preventing attenuation for strongly activated neurons. Conversely, as $z \to -\infty$,, the prime core satisfies $c(z) \to -\infty$, yielding $G(z) \to 0$. In this case, $\phi_{\text{PrimeTanh}}(z) \to 0^-$, meaning large negative activations, are smoothly attenuated rather than clipped.

4 Methodology and Experiments

Evaluation is performed in a regime dominated by severe class imbalance, focusing on the behavior of probabilistic output gates. Architectural configurations and optimization settings are held constant across all experiments, limiting sources of variability and ensuring that observed effects originate from the output nonlinearity rather than from structural or training-related adjustments.

Experiment Replicability

Reproducibility is enforced by constraining the experimental pipeline to explicitly documented operations. No external preprocessing steps, class reweighting schemes, threshold adjustments, or auxiliary optimization procedures are introduced beyond those described in the methodology. All implementation details required to replicate the experiments are made publicly available.

The complete source code, including the custom logistic regression model and the explicit definitions of all probabilistic output gates, is accessible through the authors' repository (Aguilar & Gutierrez, 2025). The dataset employed in the experiments originates from the publicly available Loan Default Prediction repository on Kaggle (Das, 2021), with the exact CSV version used retrieved directly from the corresponding GitHub source to avoid ambiguity across dataset revisions.

All experiments are implemented in Python using NumPy, Pandas, and Scikit-learn. Numerical operations rely on deterministic pseudo-randomness governed by a fixed global seed, ensuring consistent data splits and stable cross-validation behavior across repeated runs.

Dataset Characterization and Preprocessing

The dataset comprises 10,000 observations described by three numerical predictors: a binary indicator of employment status, bank balance as a continuous variable, and annual salary measured on a continuous scale.

Prior to model training, data quality was assessed to ensure that experimental outcomes would not be influenced by structural artifacts. No missing values are present across the full sample, and no exact duplicate records were identified, allowing the analysis to proceed without imputation or record filtering. The target variable exhibits a pronounced imbalance, with only 333 observations corresponding to default events against 9,667 non-default cases. This 1:29 ratio places the task well outside regimes where accuracy provides a meaningful performance signal, as a trivial majority-class predictor would exceed 96% accuracy while failing entirely at risk detection. For this reason, subsequent evaluation emphasizes recall-sensitive and correlation-based metrics.

Feature distributions display heterogeneous behavior across predictors. Bank balance presents a right-skewed distribution, with a mean of 10,024.50 and a standard deviation of 5,804.58, alongside a limited number of extreme values identified through the interquartile range criterion. Annual salary, by contrast, spans a broader numerical range, with a mean of 402,203.78 and no statistically significant outliers detected. To prevent scale-induced distortions, all features are standardized using z-score normalization performed independently within each cross-validation fold. This procedure ensures that information from the test partitions does not leak into the training process.

Cross-Validation Protocol

Model evaluation is carried out using a stratified five-fold cross-validation scheme designed to preserve the minority-class proportion across all data partitions. This stratification ensures that each training and testing split reflects the intrinsic difficulty imposed by the underlying class imbalance. Within each fold, feature standardization is performed exclusively on the training subset and subsequently applied to the corresponding test partition. Model parameters are reinitialized at every iteration, and probability estimates are computed on unseen data using the selected output gate.

To maintain strict comparability across probabilistic functions, classification decisions are derived from a fixed threshold of 0.5 throughout all experiments. Performance metrics are computed independently for each fold and reported as aggregated mean

values accompanied by their corresponding standard deviations, providing a stable estimate of both central tendency and variability.

Evaluation Metrics

Model performance is assessed through a combination of discrimination, correlation, and calibration-oriented criteria. While overall accuracy is retained for reference, greater emphasis is placed on metrics that remain informative under extreme class imbalance, including Recall, F1-score, Balanced Accuracy and the Matthews Correlation Coefficient. Complementary measures such as Precision, Cohen's Kappa, AUC, and LogLoss are incorporated to capture additional aspects of predictive behavior and probabilistic reliability. To ensure numerical stability during loss evaluation, predicted probabilities are explicitly constrained to the interval $[10^{-8}, 1\text{-}10^{-8}]$ avoiding logarithmic singularities without introducing artificial smoothing mechanisms. Confusion matrices are aggregated across cross-validation folds, allowing absolute counts of true positives, false negatives, false positives, and true negatives to be analyzed alongside averaged performance metrics.

Experiment I: Isolation of Output Gate Effect

A linear logistic classifier serves as the reference point for the initial experimental analysis. In the absence of hidden layers, the mapping from logits to probabilities collapses onto a single nonlinear transformation, leaving the output gate as the only source of nonlinearity in the decision pipeline. Under these conditions, the manner in which minority-class evidence is either attenuated or preserved is governed entirely by the shape of the output function.

Such a configuration places output gates in a particularly restrictive regime. When representational depth is removed, probability calibration and decision sensitivity rely exclusively on the behavior of the terminal nonlinearity. Saturation effects therefore manifest directly, especially under severe class imbalance, and variations in recall-oriented or correlation-based metrics can no longer be attributed to latent feature transformations or architectural expressiveness.

To prevent external corrective mechanisms from influencing this behavior, the experimental setup excludes class reweighting strategies, focal loss formulations, threshold manipulation, and all forms of data resampling. Architectural parameters are held constant across runs, further constraining variability. Within this setting, observed performance differences arise from the functional form of the output nonlinearity alone, enabling a direct and reproducible comparison between the standard Sigmoid and the proposed PrimeSigmoid.

The analysis spans the complete family of probabilistic output gates formally introduced in Section 2. Alongside the conventional Sigmoid baseline, the evaluation includes temperature-scaled Sigmoid variants with temperatures set to 0.5 and 2.0, a rescaled hyperbolic tangent, the Probit function, and the HardSigmoid formulation.

Optimization Parameters and Training Dynamics

Model optimization is carried out through gradient-based learning under a fixed configuration shared across all output gates. A learning rate of 0.05 is adopted, balancing numerical stability against convergence speed in the presence of severe class imbalance, where excessively aggressive updates tend to induce oscillatory behavior while overly conservative rates delay convergence without measurable gains. Training is allowed to proceed for up to 5,000 epochs, providing sufficient horizon for gates exhibiting slow asymptotic gradient decay, including PrimeSigmoid, to reach stable solutions.

Convergence is regulated through an early stopping criterion defined in relative terms rather than absolute loss thresholds. Training halts when the relative improvement in the loss function falls below 0.1% over a sliding window of ten epochs and remains below this threshold for twenty consecutive evaluations. This formulation is particularly relevant under imbalanced regimes, where loss values may decrease marginally even as recall-oriented metrics continue to evolve meaningfully. By relying on relative change, the stopping mechanism avoids premature termination while preventing unnecessary overtraining once functional stabilization is reached.

Experiment II: Structural Robustness in Neural Settings

Beyond the linear setting, the empirical analysis is extended to a neural configuration to examine whether the behavior observed at the output layer persists when decision boundaries are mediated by nonlinear internal representations. In this context, hidden

layers are introduced without altering the role of the output gate, which remains solely responsible for mapping the final logit to a probabilistic interpretation.

This extension allows the interaction between internal feature transformations and the terminal nonlinearity to be examined without redefining the proposed PrimeSigmoid as a hidden activation. Instead, the focus remains on its function as an output gate, while internal representations are permitted to vary. Under this setting, stability in recall-oriented and correlation-based metrics indicates that performance gains are not contingent on linear separability or architectural simplicity.

To further disentangle representational capacity from output behavior, hidden-layer activations are systematically varied while the probabilistic gate is held as an explicit experimental factor. Improvements that remain consistent across heterogeneous hidden activations can therefore be attributed to structural properties of the output nonlinearity rather than to configuration-specific synergies.

The resulting model retains a minimalist form, consisting of a single weight vector $w \in R^3$ and a scalar bias b optimized via gradient descent on the binary cross-entropy loss. This configuration preserves analytical tractability while allowing nonlinear feature transformations to influence the decision surface upstream from the output gate.

The evaluation spans the full grid of hidden activation functions formally introduced in Section 2, including the proposed PrimeTanhGate, in combination with all probabilistic output gates defined in the same section. Standard architectural baselines such as ReLU, LeakyReLU, ELU, Tanh, GELU, Swish, and Mish are included to contextualize the behavior of the proposed mechanisms within established neural design choices.

Model Architecture and Combinatorial Grid

Structural robustness is examined using a shallow multilayer perceptron designed to minimize architectural confounding while allowing controlled nonlinear transformations. The network operates on three standardized financial input features, which are processed through a single fully connected hidden layer comprising sixteen units. A single scalar logit is produced at the output and subsequently mapped to a probability through the selected output gate.

This minimalist configuration avoids depth-induced effects while remaining expressive enough to test interactions between internal representations and probabilistic output functions. To ensure comprehensive coverage, all hidden activation functions formally defined in Section 2.2, including the proposed PrimeTanhGate, are paired with every output gate introduced in Section 2. The resulting Cartesian product yields sixty-three distinct architectural combinations, each evaluated under identical training and evaluation conditions.

Training Protocol

Model training follows a uniform optimization regime applied consistently across all sixty-three architectural configurations. Parameter updates are performed using the Adam optimizer under a fixed learning rate, maintaining parity with the linear baseline established in Experiment I. No class reweighting strategies, focal loss adjustments, or resampling techniques are introduced, thereby preserving the isolation of nonlinear functional effects.

Classification decisions are derived from a constant threshold of 0.5 across all models, preventing performance variation from arising through post-hoc boundary adjustments. Predicted probabilities are constrained to the interval $[10^{-8}, 1\text{-}10^{-8}]$ prior to loss computation to avoid numerical instabilities associated with logarithmic singularities. This uniform training regime ensures that observed differences reflect the structural properties of the nonlinear transformations rather than optimizer-specific or threshold-specific artifacts.

5 Experimental Results and Analysis

Performance Analysis: Experiment I (Linear Baseline)

Empirical results are reported under a class distribution characterized by a 1:29 imbalance between default and non-default events. Under these conditions, overall accuracy remains largely invariant across models and provides limited discriminatory insight, with values clustering around 96–97%. In contrast, recall-oriented and correlation-based metrics exhibit substantial variation across probabilistic output gates, reflecting differences in minority-class sensitivity.

Comparative performance across all evaluated output gates is summarized in Table 1, with aggregated confusion matrices reported separately in Table 2 to provide absolute counts of correct and incorrect classifications. Metrics are averaged across five stratified cross-validation folds and reported alongside their corresponding standard deviations.

Table 1 reports aggregated performance metrics across five stratified folds. While the standard Sigmoid baseline achieves an accuracy of 0.9685, its recall remains limited to 0.0661, corresponding to only 22 correctly identified default cases out of 333. Temperature-scaled variants modify the curvature of the sigmoid but exhibit inconsistent effects. Increasing the temperature to 2.0 further suppresses recall, whereas a sharper configuration with T set to 0.5 partially recovers minority detection at the cost of increased variance.

The proposed PrimeSigmoid displays a distinct performance profile. Accuracy and AUC remain comparable to the baseline, while recall increases to 0.2341, yielding 78 detected defaults. This improvement is accompanied by stable precision and a marked increase in the Matthews Correlation Coefficient, indicating balanced predictive behavior rather than aggressive boundary shifting. Corresponding confusion matrices, summarized in Table 2, reflect a redistribution of false negatives toward true positives without inducing a disproportionate rise in false alarms.

Table 1. Comprehensive Performance Metrics (5-Fold Stratified CV). Mean ±Standard Deviation

Output Gate	Accuracy	Bal. Acc	Precision	Recall	MCC	AUC	LogLoss
Sigmoid (Baseline)	0.9685 ± 0.0010	0.5329 ± 0.0132	**0.8664 ± 0.1462**	0.0661 ± 0.0264	0.2291 ± 0.0501	0.9486 ± 0.0130	0.0927 ± 0.0028
PrimeSigmoid	**0.9723 ± 0.0020**	**0.6159 ± 0.0227**	0.7808 ± 0.1051	**0.2341 ± 0.0450**	**0.4171 ± 0.0646**	**0.9487 ± 0.0131**	**0.0866 ± 0.0035**
Sigmoid (T = 2.0)	0.9674 ± 0.0009	0.5120 ± 0.0078	0.8000 ± 0.4000	0.0241 ± 0.0154	0.1334 ± 0.0744	0.9485 ± 0.0129	0.1011 ± 0.0022
Sigmoid (T = 0.5)	0.9701 ± 0.0018	0.5627 ± 0.0208	0.8438 ± 0.1351	0.1262 ± 0.0415	0.3162 ± 0.0668	0.9486 ± 0.0131	0.0873 ± 0.0033
Tanh Rescaled	0.9701 ± 0.0018	0.5627 ± 0.0208	0.8438 ± 0.1351	0.1262 ± 0.0415	0.3162 ± 0.0668	0.9486 ± 0.0131	0.0873 ± 0.0033
Probit	0.9690 ± 0.0013	0.5404 ± 0.0170	0.8750 ± 0.1466	0.0812 ± 0.0341	0.2562 ± 0.0624	0.9486 ± 0.0131	0.0879 ± 0.0033
HardSigmoid	0.9667 ± 0.0002	0.5000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.9471 ± 0.0128	0.0960 ± 0.0040

Baseline Limitations Under Extreme Imbalance

Under extreme imbalance, the standard Sigmoid exhibits a strongly conservative decision profile. Although overall accuracy approaches 96.9%, minority detection remains limited. Of the 333 actual default cases, only 22 are correctly identified, while 311 remain undetected. This behavior aligns with the rapid saturation characteristic of the logistic function for negative logits, where gradients diminish sharply and minority-class contributions become attenuated during optimization.

Temperature scaling modifies the curvature of the logistic transformation but does not fundamentally alter this structural tendency. A higher temperature flattens the response curve yet further reduces recall, whereas a lower temperature sharpens the decision boundary and partially improves detection rates. These adjustments, however, operate through linear rescaling of the logit and do not change the exponential saturation mechanism.

By contrast, PrimeSigmoid preserves discrimination capacity while altering the growth dynamics of the logit transformation. Accuracy and AUC remain comparable to the baseline, while recall increases to 0.2341. Precision remains stable at 0.7808, and LogLoss decreases from 0.0927 to 0.0866, indicating improved probabilistic calibration. The Matthews Correlation Coefficient rises to 0.4171, reflecting a more balanced distribution of true positives and true negatives across folds.

The improvement in minority detection does not arise from threshold manipulation or auxiliary optimization strategies, but from a moderation of saturation dynamics induced by the logarithmic term $\ln[2 + |x|]$. This transformation delays extreme compression of the probability curve, maintaining gradient sensitivity in regions where rare-event signals are typically suppressed.

PrimeSigmoid: Recall Recovery and Probabilistic Integrity

PrimeSigmoid exhibits a distinct performance profile when compared to the standard Sigmoid baseline. Overall accuracy reaches 97.23 percent and the area under the ROC curve remains at 0.9487, while recall increases to 0.2341, corresponding to 78 correctly identified default events. Relative to the baseline, this reflects an improvement of approximately 254 percent in minority-class detection.

This increase in recall is not accompanied by degradation in complementary performance dimensions. Precision remains stable at 0.7808, indicating that gains in sensitivity do not arise from indiscriminate positive predictions. Probabilistic calibration also improves, with LogLoss decreasing from 0.0927 under the baseline configuration to 0.0866 with PrimeSigmoid, and the Matthews Correlation Coefficient rising to 0.4171. Together, these shifts indicate a more balanced distribution of correct predictions across classes rather than effects induced by aggressive threshold behavior.

The observed pattern is consistent with the structural modification introduced at the output layer. By moderating the effective growth of the logit through the logarithmic term $\ln(2 + |x|)$, PrimeSigmoid delays saturation and maintains gradient sensitivity in regions associated with rare events, allowing minority-class signals to influence training dynamics more persistently.

Table 2. Aggregated Confusion Matrices (Total Cases: 10,000).

Output Gate	TN	FP	FN	TP
Sigmoid (Baseline)	9,663	4	311	22
PrimeSigmoid	**9,645**	**22**	**255**	**78**
Sigmoid ($T = 2.0$)	9,666	1	325	8
Sigmoid ($T = 0.5$)	9,659	8	291	42
Tanh Rescaled	9,659	8	291	42
Probit	9,663	4	306	27
HardSigmoid	9,667	0	333	0

Comparison Against Alternative Gates

Other nonlinear gates show intermediate or failing behavior. The Probit function slightly improves Recall from 0.0812 relative to the baseline but remains substantially inferior to PrimeSigmoid. HardSigmoid fails completely, collapsing Recall to zero, which demonstrates that piecewise linear approximations with large zero-gradient regions are unsuitable for imbalanced regimes. Across all evaluated functions, PrimeSigmoid is the only gate that simultaneously improves Recall, F1-score, and MCC without degrading AUC.

Reliability Diagram

Probabilistic calibration is examined to assess whether gains in minority-class detection are accompanied by distortions in probability estimates. In addition to discrimination-oriented metrics, calibration quality is quantified using the Brier score and the Expected Calibration Error (ECE), providing complementary views of probabilistic accuracy and reliability.

Reliability diagrams reveal systematic differences in confidence behavior across output gates. The standard Sigmoid exhibits increasing overconfidence in the mid-to-high probability range, deviating progressively from the identity mapping. In contrast, the proposed PrimeSigmoid remains closer to the diagonal across a broad interval, particularly between probability values of approximately 0.4 and 0.85, indicating improved alignment between predicted probabilities and observed frequencies.

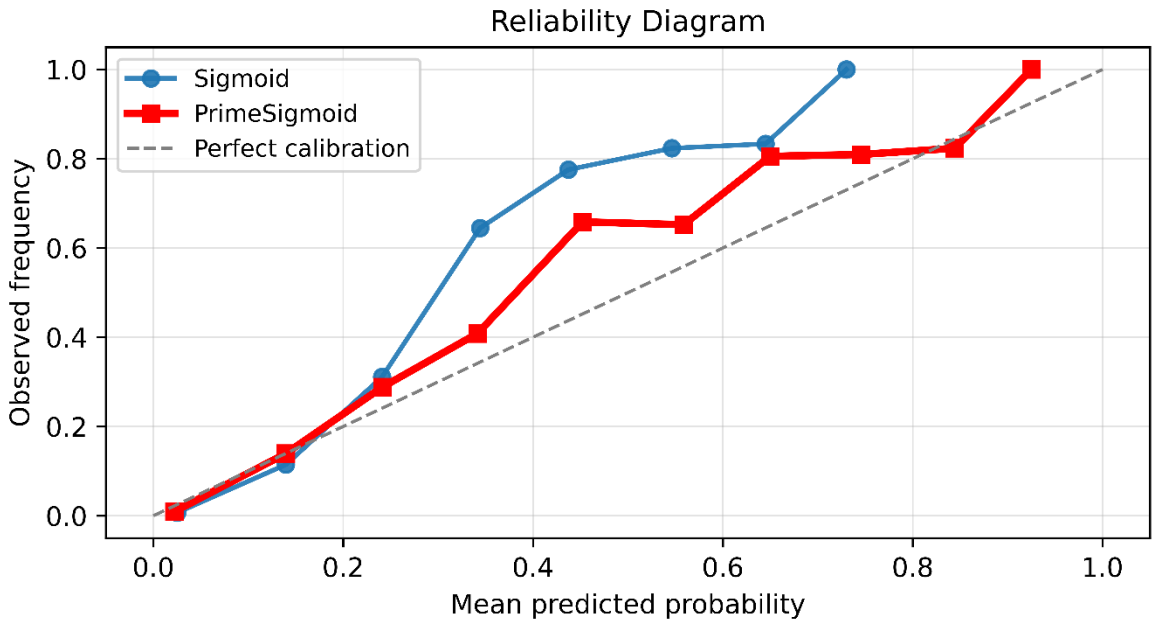

Fig. 2. Reliability Diagram

Quantitative calibration metrics corroborate this pattern. PrimeSigmoid achieves a lower Brier score (0.0219) relative to the standard Sigmoid (0.0235), alongside a substantial reduction in ECE from 0.0245 to 0.0149. These results indicate that the moderation of saturation dynamics introduced at the output layer does not compromise probabilistic integrity and, instead, yields more reliable confidence estimates under severe class imbalance.

Financial Risk Implications

From the perspective of financial risk modeling, the observed differences carry direct operational implications. In lending environments, the cost associated with a missed default typically exceeds that of a false alarm, introducing an inherent asymmetry in error tolerance. Under this asymmetry, output functions that systematically suppress minority-class signals effectively encode a prior favoring majority-class stability.

PrimeSigmoid alters this implicit bias without modifying the loss function, class weights, or decision threshold. Such a mechanism is particularly relevant in regulated settings where modifications to training objectives or decision policies may be constrained. Adjustments at the level of the output gate preserve the overall modeling framework while enabling sensitivity to rare but economically significant events to be recovered through structural means.

Performance Analysis: Experiment II (MLP Configurations)

Results from the exhaustive grid search spanning sixty-three architectural configurations are evaluated under the same imbalance regime as Experiment I. Given the limited informativeness of accuracy in this context, models are ranked primarily according to F1-score, with Matthews Correlation Coefficient and Balanced Accuracy used as secondary criteria. The five highest-ranked configurations are retained for closer inspection.

Across the global ranking, configurations employing PrimeSigmoid at the output layer recur consistently among the top-performing models, despite substantial variation in hidden-layer activations. Accuracy remains tightly clustered around 97.3% across all evaluated networks, whereas recall and F1-score exhibit meaningful dispersion, indicating differential sensitivity to minority-class events. Table 3 summarizes the performance metrics associated with the highest ranked configurations.

Table 3. Top 5 Global Configurations (Ranked by F1-score). Mean ± Standard Deviation

Rank	Hidden	Output Gate	F1-score	Recall	Precision	MCC	Bal. Acc
1	ELU	**PrimeSigmoid**	**0.4622 ± 0.0651**	**0.3453 ± 0.0568**	0.7058 ± 0.0875	**0.4815 ± 0.0664**	**0.6702 ± 0.0286**
2	ReLU	PrimeSigmoid	0.4573 ± 0.0639	0.3452 ± 0.0572	0.6837 ± 0.0760	0.4732 ± 0.0634	0.6699 ± 0.0288
3	Tanh	Sigmoid (T = 2.0)	0.4571 ± 0.0634	0.3362 ± 0.0511	**0.7162 ± 0.0857**	0.4794 ± 0.0661	0.6658 ± 0.0260
4	Tanh	PrimeSigmoid	0.4561 ± 0.0628	0.3362 ± 0.0511	0.7120 ± 0.0852	0.4778 ± 0.0653	0.6658 ± 0.0260
5	LeakyReLU	PrimeSigmoid	0.4553 ± 0.0633	0.3362 ± 0.0511	0.7083 ± 0.0887	0.4764 ± 0.0663	0.6657 ± 0.0260

The leading model combines an ELU hidden activation with a PrimeSigmoid output gate, achieving the highest balance between recall and precision, as reflected in its Matthews Correlation Coefficient. Comparable performance is observed when PrimeSigmoid is paired with ReLU or Tanh activations, suggesting that the contribution of the output gate is not contingent on a specific internal representation. The only non-PrimeSigmoid configuration appearing in the top tier relies on temperature-scaled Sigmoid behavior, with performance driven by parameter tuning rather than by structural modification of the output nonlinearity.

The behavior of the proposed PrimeTanhGate is examined in Table 4 by situating its performance within the broader landscape of hidden activation functions evaluated in Experiment II. When combined with PrimeSigmoid at the output layer, PrimeTanhGate yields stable performance across recall, F1-score, and correlation-based metrics, remaining within the range observed for established activations such as ReLU, ELU, and Mish.

Table 4. Comparative Analysis: PrimeTanhGate (Best Config) vs. Global Top-5 Average.

Metric / Aspect	PrimeTanhGate (Best)	Global Top-5 (Avg)
Output Gate Strategy	**PrimeSigmoid**	**PrimeSigmoid (4/5)**
Max Recall	≈ 0.330	≈ 0.345
Max F1-score	≈ 0.444	≈ 0.462
Max MCC	≈ 0.462	≈ 0.482

The strongest configuration involving this activation pairs PrimeTanhGate with PrimeSigmoid, achieving a Balanced Accuracy of 0.6624, a Recall of 0.3301, and an MCC of 0.4621. These values are comparable to those attained by the highest-ranked architectures in the global grid, as summarized in Table 4, indicating that the proposed hidden activation integrates coherently within the evaluated neural setting.

When contrasted with the average performance of the global top-ranked configurations, PrimeTanhGate exhibits slightly lower peak values across recall and MCC, while maintaining consistent behavior across folds. This positioning suggests that, within the tested architecture, gains in minority-class sensitivity are primarily governed by the choice of output nonlinearity rather than by the specific form of the hidden activation. The logarithmic moderation introduced at the output layer appears sufficient to recover sensitivity to rare events across a range of internal representations, whether classical or prime-inspired.

6 Considerations

The methodological design of this study operates under explicit constraints intended to preserve interpretability and causal attribution. External imbalance-handling mechanisms such as class reweighting, focal loss formulations, resampling procedures, and threshold adjustments are intentionally excluded. Although widely used in practice, these techniques intervene at different

stages of the learning pipeline by altering the optimization objective or post-training decision boundary. Their omission ensures that observed performance differences arise from the structural properties of the output activation itself. Reported improvements should therefore be interpreted as conservative estimates of the functional contribution of the proposed transformation.

Empirical evaluation is conducted on a single real-world financial risk dataset characterized by severe class imbalance. While representative of high-stakes decision environments, validation across additional domains, including fraud detection and medical diagnosis, would further clarify the extent to which the observed behavior generalizes beyond credit-risk modeling contexts.

Architectural complexity is deliberately constrained to shallow configurations in order to prevent representational depth from masking the influence of the output nonlinearity. The behavior of logarithmically moderated gates within deeper residual or attention-based architectures remains an open empirical question and a natural direction for subsequent investigation.

The introduction of PrimeTanhGate is positioned as an exploratory structural extension rather than as a definitive substitute for established activations such as Mish or Swish. Its role within this study is to demonstrate compatibility with PrimeSigmoid and to assess whether prime-inspired transformations integrate coherently within standard neural settings.

The isolation of the output activation reflects a conceptual distinction between structural probability mapping and optimization-based imbalance correction. Techniques such as focal loss or class weighting modify gradient allocation during training, while threshold tuning operates at the decision phase. By contrast, PrimeSigmoid reshapes the saturation dynamics of the probability mapping itself, moderating compression in regions where minority-class signals are typically attenuated. This distinction becomes particularly relevant in financial risk modeling environments subject to governance constraints, where decision thresholds are often fixed by policy and regulatory requirements. Under such conditions, adjustments at the output mapping level provide an alternative mechanism for enhancing minority detection while preserving procedural stability, transparency, and reproducibility across reporting periods.

7 Conclusions

The analysis presented in this work explores an alternative treatment of output-layer saturation under extreme class imbalance by modifying the probabilistic mapping itself. Rather than relying on linear growth followed by exponential compression, the proposed PrimeSigmoid introduces a logarithmic moderation inspired by prime number asymptotics, reshaping the saturation dynamics governing probability formation.

From a theoretical standpoint, this modification alters how gradient sensitivity decays for large-magnitude pre-activations. By slowing saturation through inverse logarithmic normalization, sensitivity to rare-event signals is retained in regions where standard sigmoidal functions typically collapse. This structural adjustment directly targets a well-known failure mode of conventional output gates, namely the systematic attenuation of minority-class contributions during training.

Empirical evaluation across linear classifiers and a comprehensive grid of multilayer perceptron configurations reflects this shift in behavior. Improvements in Recall, F1-score, and Matthews Correlation Coefficient emerge consistently, while probabilistic calibration, as measured by AUC and LogLoss, remains stable. Notably, these effects arise without altering the loss formulation, introducing class-dependent weights, or adjusting decision thresholds, isolating the contribution of the output nonlinearity itself.

Results from the extended grid analysis further indicate that the observed behavior is not confined to a specific architectural configuration. PrimeSigmoid appears recurrently among the highest-ranked models across heterogeneous hidden representations, suggesting that its effect is structurally anchored in the output mapping rather than contingent on internal feature transformations. The exploratory PrimeTanhGate extends this perspective by demonstrating that prime-inspired moderation can be incorporated into hidden activations while maintaining stable gradients and competitive performance, although the dominant contribution remains concentrated at the output layer.

Taken together, these findings position logarithmically moderated nonlinearities as a viable structural alternative for classification tasks dominated by extreme imbalance. By intervening at the level of saturation dynamics rather than through optimization heuristics or post-hoc decision adjustments, PrimeSigmoid offers a complementary mechanism for recovering minority-class sensitivity. This approach aligns naturally with high-stakes domains such as financial risk modeling, where procedural stability, calibration integrity, and regulatory constraints often limit the scope for threshold tuning or loss reengineering.

References

Aguilar, J., & Gutiérrez, M. (2025). *PrimeSigmoid and PrimeTanhGate implementation* [Repositorio de software]. GitHub. https://github.com/derivado29/Primesigmoid_PrimeTanhGate_Implementation

Bliss, C. I. (1934). The method of probits. *Science, 79*(2037), 38–39. https://doi.org/10.1126/science.79.2037.38

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), Article 6. https://doi.org/10.1186/s12864-019-6413-7

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (ELUs). In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1511.07289

Courbariaux, M., Bengio, Y., & David, J.-P. (2015). BinaryConnect: Training deep neural networks with binary weights during propagations. *Advances in Neural Information Processing Systems, 28*. https://arxiv.org/abs/1511.00363

Das, K. (2021). *Loan default prediction dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/kmldas/loan-default-prediction

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*. https://www.ijcai.org/Proceedings/01/Papers/161.pdf

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (*Proceedings of Machine Learning Research*, Vol. 9, pp. 249–256). https://proceedings.mlr.press/v9/glorot10a.html

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*. https://arxiv.org/abs/1706.04599

Hardy, G. H., & Wright, E. M. (1979). *An introduction to the theory of numbers* (5th ed.). Oxford University Press.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv*. https://arxiv.org/abs/1606.08415

LeCun, Y. A., Bottou, L., Orr, G. B., & Müller, K.-R. (2012). Efficient backprop. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade* (2nd ed., pp. 9–48). Springer. https://doi.org/10.1007/978-3-642-35289-8_3

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2999–3007). https://doi.org/10.1109/ICCV.2017.324

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning*. https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf

Misra, D. (2019). Mish: A self-regularized non-monotonic neural activation function. *arXiv*. https://arxiv.org/abs/1908.08681

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv*. https://arxiv.org/abs/1710.05941

Riemann, B. (1859). Ueber die Anzahl der Primzahlen unter einer gegebenen Grösse. *Monatsberichte der Berliner Akademie der Wissenschaften zu Berlin*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et physique, 10*, 113–121.