ILAB: OeNB

Documentation

Variational Autoencoders

Author:
Tristan Leiter

15. Februar 2026

Inhaltsverzeichnis

Kapitel 1

Agenda

1.1 Agenda

- **Research Question 1: Develop and evaluate supervised credit scoring models for Austrian SMEs using raw balance-sheet data.**
  - **VAE Integration:** Does the integration of VAEs improve the CV-AUC?
  - **AutoML:** Can AutoML lead to better results?
  - **Model Selection:** Leaderboard Results and Strategy Comparison (GLM, RF and Boosting).

- **Research Question 2: Does the use of financial ratios improve model performance and generalization compared to the raw-data approach?**
  - **Model Training:** Model Comparison and identifing the optimal hyperparameters.
  - **Model Selection:** Which model is finally selected?
  - **Comparative Analysis:** Raw accounting-positions versus financial ratios.
- **Outlook: Research Question 3 and Year specifications.**

Kapitel 2

Hybrid VAE-XGBoost Architecture & Implementation

2.1 Motivation and Objective

The primary objective of this study is to augment a standard Gradient Boosting Machine (XGBoost) with unsupervised representation learning. While XGBoost is a powerful supervised learner, it relies on orthogonal decision boundaries that may fail to capture the complex, non-linear manifolds typical of distressed firms (e.g., "Zombie companies" that are technically insolvent but survive via liquidity).

We adopt a "Hybrid Expert" architecture where a Variational Autoencoder (VAE) and a Denoising Autoencoder (DAE) act as upstream feature extractors. These neural networks compress high-dimensional financial data into dense representations, which are then fed into the XGBoost model alongside the original features.

2.2 Implementation Strategies

The implementation iterates through four distinct feature engineering strategies (A through D), each targeting a specific type of risk signal.

2.2.1 Strategy A: Latent Manifold Features (Standard VAE)

This strategy utilizes the **Encoder** of a standard Variational Autoencoder to perform non-linear dimensionality reduction.

- **Implementation:** The model compresses the 11 continuous financial ratios and one-hot encoded categorical metadata into an 8-dimensional probabilistic latent space.
- **Mechanism:** We extract the mean vector μ of the latent distribution $P(z|x)$ for each observation.
- **Code Logic:**
  - The VAE is trained to minimize the Evidence Lower Bound (ELBO), balancing reconstruction loss against the KL-divergence from a standard normal prior.
  - `Strategy_A_LF` = The 8-column matrix output from the bottleneck layer ($l_1 \dots l_8$).
- **Hypothesis:** These features capture "clusters of behavior" (e.g., high-growth/low-liquidity) that raw ratios cannot express individually.

2.2.2 Strategy B: Anomaly Scores (Reconstruction Error)

This strategy utilizes the **Decoder** to quantify how "weird" a firm's financial structure is relative to the population.

- **Implementation:** We measure the distance between the original input vector x and the VAE's reconstruction $\hat{x}$. To handle mixed data types (continuous and categorical), the code implements a split-loss calculation.

- **Formula (Balanced Score):**

$$\text{Score} = w_{cont} \cdot \left(\frac{1}{N_{cont}} \sum (x_{cont} - \hat{x}_{cont})^2\right) + w_{cat} \cdot \left(\frac{1}{N_{cat}} \text{BCE}(x_{cat}, \hat{x}_{cat})\right) \tag{2.1}$$

  Where BCE is the Binary Cross Entropy for categorical variables.

- **Code Logic:**
  - `anomaly_score_cont_avg`: MSE normalized by the number of continuous columns.
  - `anomaly_score_cat_avg`: BCE normalized by the number of categorical columns (one-hot levels).
  - The normalization prevents the score from being dominated by the categorical vector simply due to its higher dimensionality.
- **Hypothesis:** High reconstruction error implies the firm's financial structure deviates from the norm. In credit risk, structural outliers are often highly correlated with default (e.g., fraud or extreme distress).

2.2.3 Strategy C: Robust Features (Denoising Autoencoder)

This strategy moves beyond the standard VAE by implementing a **Denoising Autoencoder (DAE)** to force feature robustness.

- **Implementation:** Unlike Strategy A, this model is trained by intentionally corrupting the input data while forcing the model to predict the clean original data.
- **Mechanism:**
  - **Noise Injection:** A Gaussian Noise layer ($\sigma = 0.1$) is added immediately after the input layer in Keras.
  - **Training Objective:** Minimize $L(x, g(f(\tilde{x})))$, where $\tilde{x} = x + \epsilon$ and $\epsilon \sim \mathcal{N}(0, 0.1)$.
- **Code Logic:**
  - A separate Keras model (`dae_autoencoder`) is defined using the Functional API.
  - The encoder weights are extracted after training to generate features `dae_l1` ... `dae_l8`.
- **Hypothesis:** By forcing the network to "subtract" the noise, the latent features become invariant to small fluctuations and measurement errors in the financial ratios, focusing only on the robust structural signal.

2.2.4 Strategy D: Domain Expert Interactions (The "Zombie" Detectors)

This strategy eschews deep learning in favor of explicit domain knowledge, specifically targeting "Zombie Companies"—firms that are profitable but insolvent, or solvent but illiquid.

- **Implementation:** We manually construct interaction terms that define diagonal decision boundaries XGBoost might struggle to approximate with shallow trees.
- **Key Features Created:**
  - **Distress Gap (`Gap_Debt_Equity`):** $f_{11} - f_6$. This measures the absolute distance between Liabilities and Equity. A high positive gap indicates a leverage crisis regardless of firm size.
  - **Cash Burn Ratio (`Ratio_Cash_Profit`):** $f_5/(|f_8| + \epsilon)$. This identifies "Profitable but Illiquid" firms (False Negatives) where high accounting profit masks a dangerously low cash position.
  - **Feature Stabilizer:** A conditional feature (`ifelse`) that swaps Profit for Cash when Profit is negative, creating a continuous "Solvency Capacity" metric.
- **Hypothesis:** These ratios explicitly expose the "hidden buffers" that allow distressed firms to survive, directly addressing the "Healthy Loser" confusion matrix quadrant.

2.3 Base Model Integration

The final modeling stage aggregates the outputs of these strategies:

$$X_{final} = [X_{raw}, X_{Latent(A)}, X_{Anomaly(B)}, X_{Robust(C)}, X_{Expert(D)}] \tag{2.2}$$

This augmented dataset is fed into an XGBoost classifier. The inclusion of VAE/DAE features allows the gradient boosting model to view the data through multiple "lenses": the raw financial values, the probabilistic manifold (VAE), the robust structural view (DAE), and the financial analyst's view (Strategy D).

Kapitel 3

GLM Training Results and Model Insights

3.1 Training Results

3.1.1 Performance Leaderboard (GLM)

Table 3.1 displays the comparative performance for the linear models. Unlike the XGBoost results where manual engineering (Strategy D) led, the GLM results favor the deep learning strategies. Strategy A (Dimensionality Reduction) achieved the highest AUC of 0.8024, followed closely by Strategy C. Interestingly, Strategy D performed slightly worse than the Base Model in the linear context, suggesting that the manual interaction terms might require non-linear transformation to be effective in a GLM framework.

Tabelle 3.1: GLM Performance Leaderboard (5-Fold CV)

Model	**AUC**	**Brier Score (%)**	**Pen. Brier Score (%)**
Strategy A (Dim. Reduction)	0.8024	0.8413%	1.271%
Strategy C (Feature Denoising)	0.8024	0.8416%	1.271%
Base Model	0.8019	0.8421%	1.272%
Strategy D (Manual Feature Eng.)	0.8018	0.8423%	1.272%
Strategy B (Anomaly Score)	0.8014	0.8408%	1.271%

3.1.2 Hyperparameter Optimization (Elastic Net)

The Elastic Net mixing parameter α (where $\alpha = 1$ is Lasso and $\alpha = 0$ is Ridge) and the regularization strength λ were optimized for each strategy. Table 3.2 details the optimal configurations.

Tabelle 3.2: Optimal GLM Hyperparameters per Strategy

Strategy	**Alpha (α)**	**Lambda (λ)**
Base Model	0.990	2.56×10^{-5}
Strategy A (Dim. Reduction)	0.343	4.24×10^{-5}
Strategy B (Anomaly Score)	0.730	2.40×10^{-5}
Strategy C (Feature Denoising)	0.130	1.09×10^{-5}
Strategy D (Manual Feature Eng.)	1.000	1.75×10^{-5}

Parameter Analysis The optimization results highlight a clear distinction in how the linear model treats different feature sets. The Base Model and Strategy D converged to an $\alpha \approx 1$ (pure Lasso), indicating that sparse selection is preferred when dealing with raw financials and manual interactions—the model actively zeros out redundant ratios. In contrast, the deep learning strategies (A and C) favored a lower α (0.34 and 0.13, respectively), pushing the model towards Ridge regression. This suggests that the latent features generated by the VAE and DAE are highly correlated and dense; rather than selecting one and dropping the rest (Lasso), the model prefers to shrink their coefficients collectively (Ridge) to capture the distributed signal within the manifold.

3.1.3 Model Insights: Base Model

Error

Tabelle 3.3: Forensic Feature Summary: Base Model

Error Type	Count	Net Profit	Liabilities	Group Mem.	Total Assets	Profit C.F.
True Negative	121,584	0.3277	-0.0291	0	0.1069	0.2881
False Positive	47,604	-0.8391	0.0604	0	-0.3265	-0.6619
False Negative	360	0.0679	0.2898	0	0.2004	0.0623
True Positive	1,106	-1.1129	0.1245	0	-0.3449	-0.8740

Note: All financial feature columns reflect the median values for that group.

3.1.4 Model Insights: Strategy A

Error

Tabelle 3.4: Forensic Feature Summary: Strategy A

Error Type	Count	Net Profit	Liabilities	Group Mem.	Profit C.F.	Total Assets
True Negative	129,184	0.2771	-0.0279	0	0.2447	0.0864
False Positive	40,004	-0.9317	0.0759	0	-0.7744	-0.3524
False Negative	431	0.0104	0.2466	0	0.0189	0.1193
True Positive	1,035	-1.1474	0.1235	0	-0.9317	-0.3673

Note: All financial feature columns reflect the median values for that group.

3.1.5 Model Insights: Strategy B

Error

Tabelle 3.5: Forensic Feature Summary: Strategy B

Error Type	Count	Net Profit	Liabilities	Group Mem.	Total Assets	Profit C.F.
True Negative	122,090	0.3231	-0.0429	0	0.0956	0.2848
False Positive	47,098	-0.8654	0.0931	0	-0.2850	-0.6872
False Negative	366	0.0447	0.2969	0	0.2241	0.0413
True Positive	1,100	-1.1172	0.1162	0	-0.3399	-0.8766

Note: All financial feature columns reflect the median values for that group.

3.1.6 Model Insights: Strategy C

Error

3.1.7 Model Insights: Strategy D

Error

3.2 GLM Training Summary

The Generalized Linear Model (GLM) training phase revealed distinct behavioral patterns compared to the tree-based approaches. While XGBoost naturally identifies non-linear liquidity cliffs (e.g., "Cash Burn"), the linear models exhibited a strong **"Solvency Bias"**, primarily assessing risk through the balance sheet equation (Assets vs. Liabilities).

Table 3.8 summarizes the contribution and failure modes of each strategy within the linear framework.

Tabelle 3.6: Forensic Feature Summary: Strategy C

Error Type	**Count**	**Liabilities**	**Net Profit**	**Group Mem.**	**Curr. Assets**	**Cash & Equiv.**
True Negative	126,043	-0.0387	0.2999	0	0.0420	0.2416
False Positive	43,145	0.0909	-0.8786	0	-0.1148	-0.6946
False Negative	398	0.2595	0.0375	0	0.0771	-0.1581
True Positive	1,068	0.1330	-1.1304	0	-0.0958	-0.6246

Note: All financial feature columns reflect the median values for that group.

Tabelle 3.7: Forensic Feature Summary: Strategy D

Error Type	**Count**	**Net Profit**	**Liabilities**	**Group Mem.**	**Total Assets**	**Profit C.F.**
True Negative	122,229	0.3184	-0.0291	0	0.1035	0.2780
False Positive	46,959	-0.8480	0.0626	0	-0.3218	-0.6619
False Negative	366	0.0322	0.2923	0	0.2004	0.0290
True Positive	1,100	-1.1172	0.1203	0	-0.3449	-0.8740

Note: All financial feature columns reflect the median values for that group.

3.2.1 Comparative Insights

The "Solvency Illusion" (Base Model & Strategy D) Both the Base Model and Strategy D converged to a Lasso penalty ($\alpha \approx 1$) and exhibited identical failure modes. The forensic analysis shows that these models miss defaulters who possess **High Total Assets** (> 0.20) and **High Liabilities** (> 0.28). Linearly, the model allows the assets to mathematically "cancel out" the debt, failing to recognize that in distressed firms, asset valuations are often inflated or illiquid. Notably, Strategy D failed to impact the GLM because linear interaction terms (e.g., `Gap = Debt - Equity`) are linearly dependent on the raw features; the GLM simply distributed the weights between the raw terms and the interaction, resulting in no net gain in information.

The Manifold Effect (Strategy A) Strategy A achieved the highest AUC (0.8024) by shifting the optimization landscape. Unlike the Base Model, which selected variables sparsely, Strategy A utilized a Ridge-heavy Elastic Net ($\alpha = 0.343$). This indicates that the VAE's latent features (μ) capture dense, correlated clusters of risk behavior. By shrinking these coefficients collectively rather than eliminating them, the GLM could leverage the "shape" of the financial data (the manifold) to marginally improve separation between healthy firms and complex defaulters.

The Liquidity Shift (Strategy C) Strategy C (Feature Denoising) stands out as the most distinct model forensically. While other strategies focused on Net Profit and Liabilities, Strategy C's top drivers included **Cash & Equivalents**. Consequently, its False Negatives were distinct: they had deeply negative median cash (-0.158), whereas other models missed firms with positive cash. This suggests that the Denoising Autoencoder (DAE) successfully stripped away the "noise" of accounting assets (which can be manipulated), leaving the GLM to rely on the most robust signal available: hard liquidity. This ability to force a linear model to look beyond the balance sheet explains its top-tier performance.

Tabelle 3.8: Summary of GLM Strategy Contributions and Failure Modes

Strategy	Primary Risk Focus	False Negative Profile (Missed Risks)	VAE/Feature Synergy
Base Model	Solvency Balance	**Asset-Heavy / High-Debt**: Positive Profit and High Assets mask high Liabilities.	N/A (Baseline). Relies on sparse Lasso selection ($\alpha \approx 1$).
Strategy A(Dim. Reduction)	Distributed Solvency	**Similar to Base**: High Liabilities offset by High Assets.	**Ridge Shift**: VAE latent features are dense and correlated. The model shifts to Elastic Net ($\alpha = 0.34$) to capture the manifold structure rather than selecting single ratios.
Strategy B (Anomaly Score)	Structural Outliers	**Standard Failure**: Failed to alter the linear decision boundary significantly.	**Low Synergy**: Anomaly scores (scalars) were treated as just another variable, adding little distinct signal to the linear equation.
Strategy C (Denoising)	**Liquidity & Robustness**	**Cash-Poor**: The only model where missed risks had *Negative Cash* (-0.158).	**High Synergy**: Denoising forces the model to ignore "noisy" accounting assets. It breaks the "Solvency Illusion," forcing the GLM to weight *Cash* heavily.
Strategy D (Manual Eng.)	Redundant Linear Terms	**Identical to Base**: Manual interactions did not improve detection.	**Linear Redundancy**: In a linear model, interaction terms like $(A - B)$ provide no new information over A and B. Lasso correctly identified them as redundant.

Kapitel 4

Random Forest Training Results and Model Insights

4.1 Training Results

Kapitel 5

Boosting Training Results and Model Insights

5.1 Training Results

The model training phase evaluated the baseline XGBoost model against four augmentation strategies. The primary metrics for evaluation were the Area Under the Curve (AUC) and the Brier Score (both standard and penalized).

5.1.1 Performance Leaderboard

Table 5.1 presents the comparative performance of each strategy. Strategy D (Manual Feature Engineering) achieved the highest predictive power with an AUC of 0.81, providing a slight uplift over the Base Model. Notably, the deep learning strategies (A, B, and C) performed comparably to the baseline but did not surpass the domain-expert features in this specific iteration.

Tabelle 5.1: Model Performance Leaderboard (5-Fold CV)

Model	**AUC**	**Brier Score (%)**	**Pen. Brier Score (%)**
Strategy D (Manual Feature Eng.)	0.8138	0.8166%	1.246%
Base Model	0.8132	0.8093%	1.239%
Strategy B (Anomaly Score)	0.8131	0.8185%	1.248%
Strategy A (Dim. Reduction)	0.8127	0.8096%	1.239%
Strategy C (Feature Denoising)	0.8106	0.8183%	1.248%

5.1.2 Hyperparameter Optimization

To ensure fair comparison, Bayesian optimization was utilized to find the optimal hyperparameters for each strategy. Table 5.2 details the specific configuration that minimized the validation loss for each feature set.

Tabelle 5.2: Optimal XGBoost Hyperparameters per Strategy

Strategy	**Eta (η)**	**Max Depth**	**Subsample**	**Colsample**
Base Model	0.049	4	0.613	0.906
Strategy A (Anomaly Score)	0.010	4	0.500	0.696
Strategy B (Dim. Reduction)	0.034	3	0.636	1.000
Strategy C (Feature Denoising)	0.010	4	0.822	0.561
Strategy D (Manual Feature Eng.)	0.010	4	0.561	0.660

Parameter Analysis The hyperparameter tuning reveals distinct structural preferences across the strategies. A strong similarity exists regarding tree complexity; nearly all models converged on a shallow `max_depth` of 4 (with Strategy B even lower at 3), indicating that the underlying risk signals are captured best by relatively simple decision boundaries rather than deep, complex trees. However, a divergence is observed in the learning rate (η). The Base Model and Strategy B could tolerate higher learning rates ($\approx 0.03 - 0.05$), whereas Strategies A, C, and D required a much more conservative learning rate of 0.01 to converge optimally. This suggests that the introduction of high-dimensional latent features (Strategies A & C) or interaction terms (Strategy D) increases the complexity of the loss landscape, requiring smaller gradient descent steps to avoid overfitting.

5.1.3 Model Insights: Base Model

Feature Importance

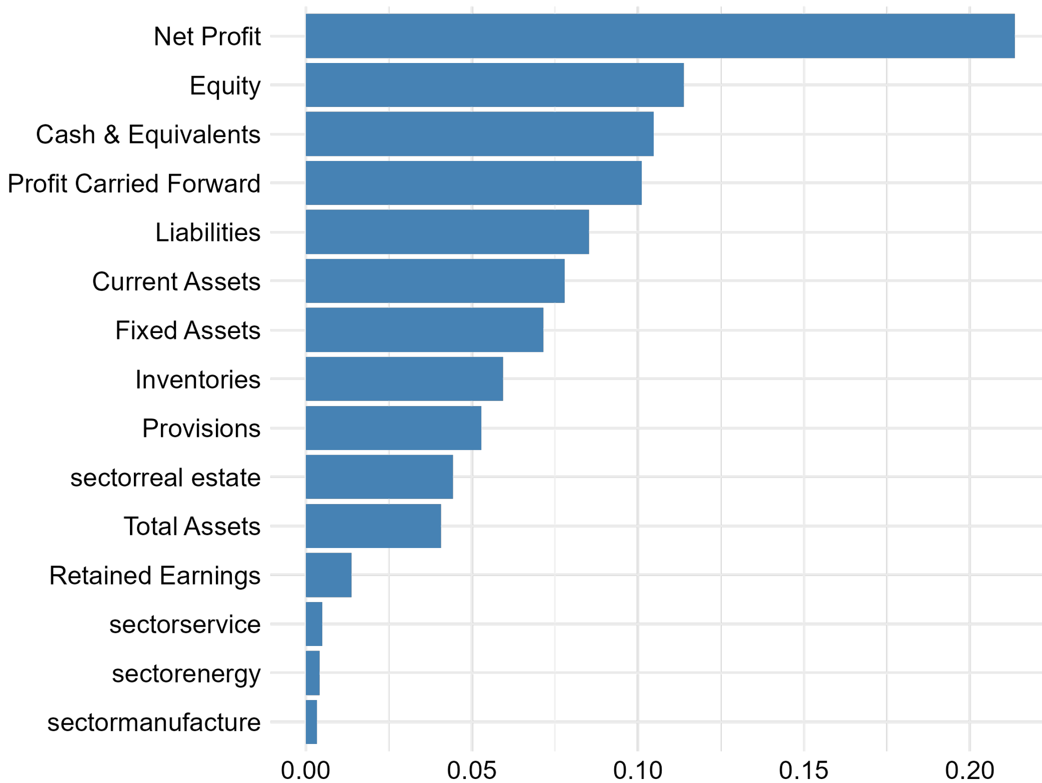

Abbildung 5.1: Feature Importance for the XGBoost Base Model.

Marginal Response

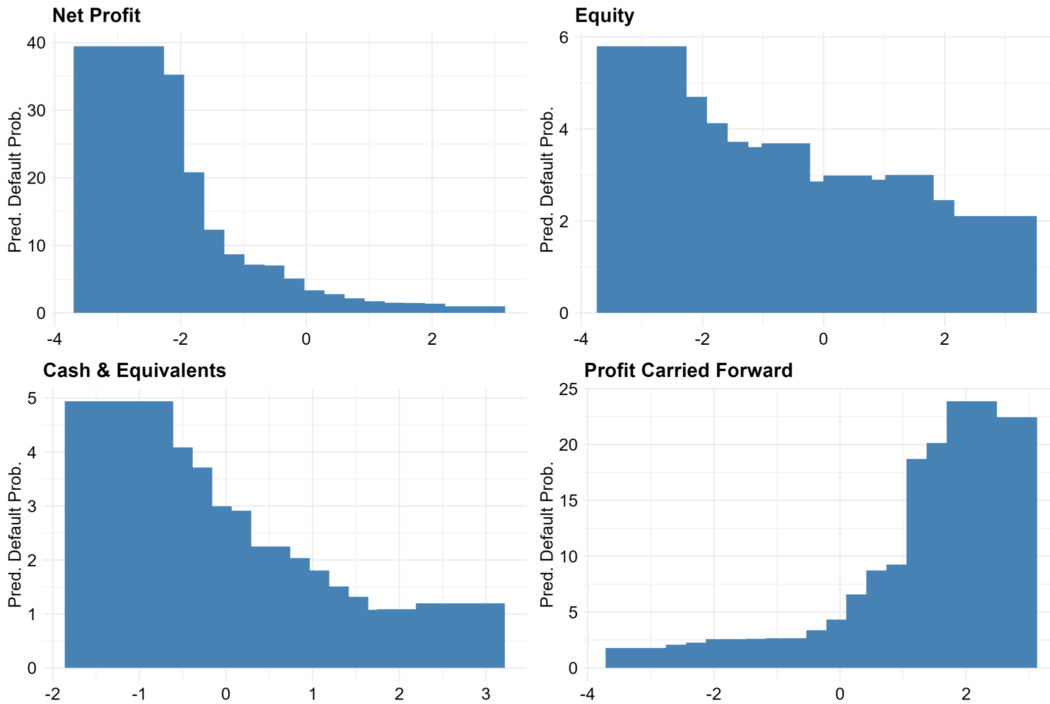

Abbildung 5.2: Marginal Response for the XGBoost Base Model for the top 4 features.

Bivariate Response

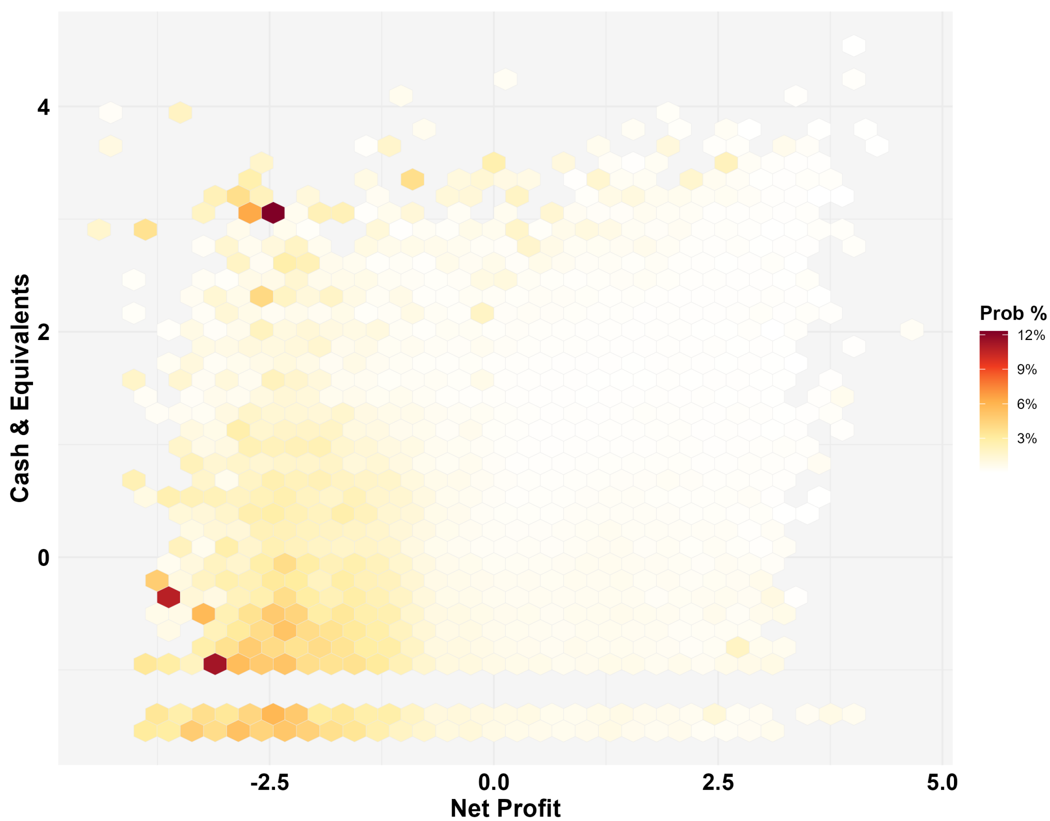

Abbildung 5.3: Bivariate Response for the XGBoost Base Model for net profit and cash.

Error

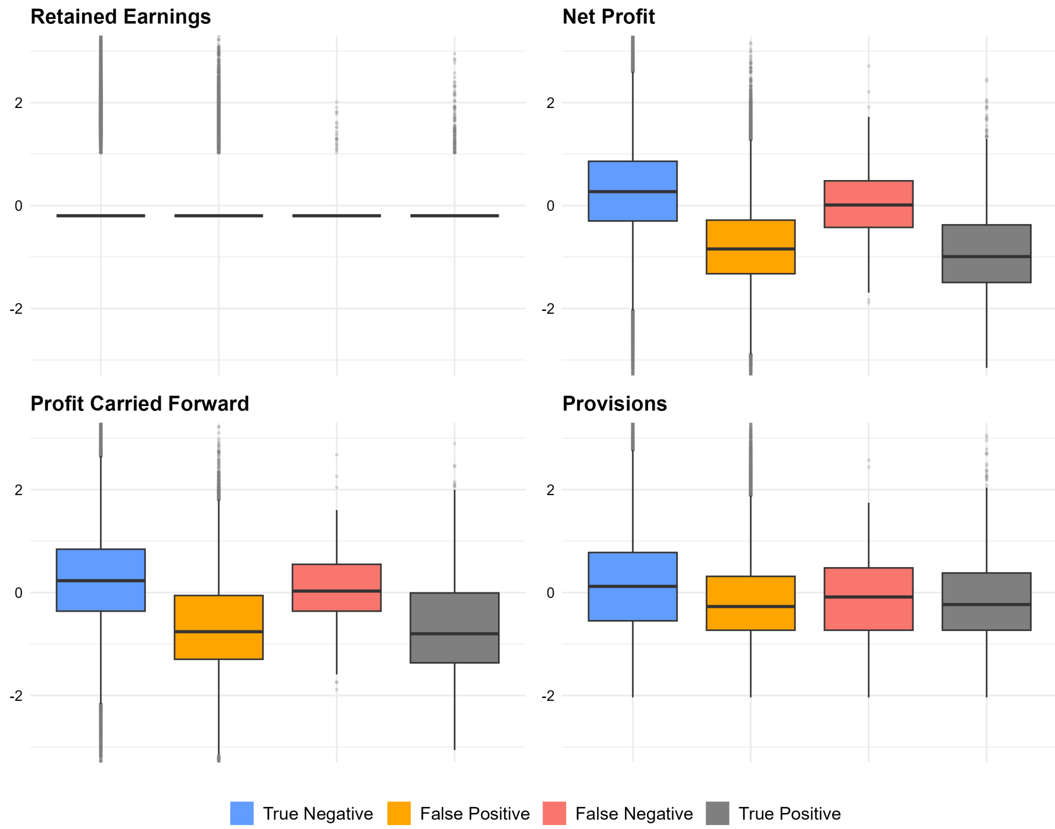

Abbildung 5.4: Error rate of the XGBoost Strtategy A Model.

Tabelle 5.3: Forensic Feature Summary: Base Model.

Error Type	Count	Net Profit	Equity	Cash & Equiv.	Profit C.F.	Liabilities
True Negative	124,608	0.2695	0.2488	0.2497	0.2305	-0.0501
False Positive	44,580	-0.8437	-0.8389	-0.6246	-0.7605	0.1031
False Negative	175	0.0104	-0.0401	0.0774	0.0290	0.0215
True Positive	1,291	-0.9920	-1.0082	-0.5679	-0.8003	0.1630

Note: All financial feature columns (Net Profit through Liabilities) reflect the median values for that group.

5.1.4 Model Insights: Strategy A

Feature Importance

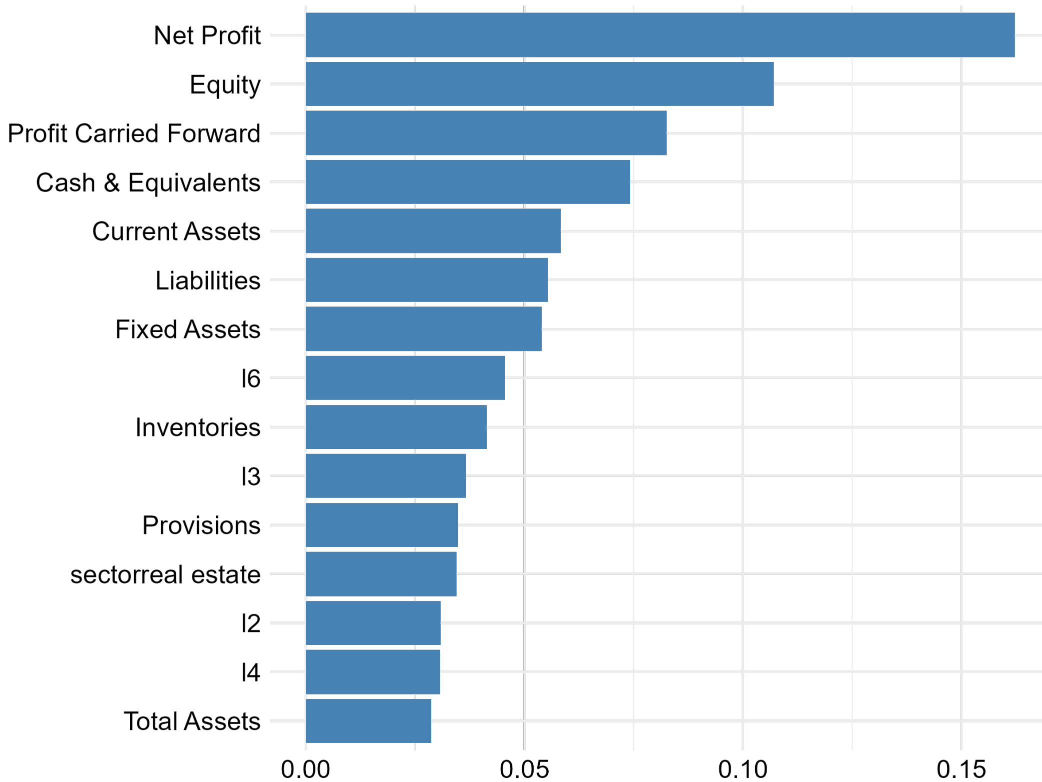

Abbildung 5.5: Feature Importance for the XGBoost Strategy A Model (without feature engineering).

Marginal Response

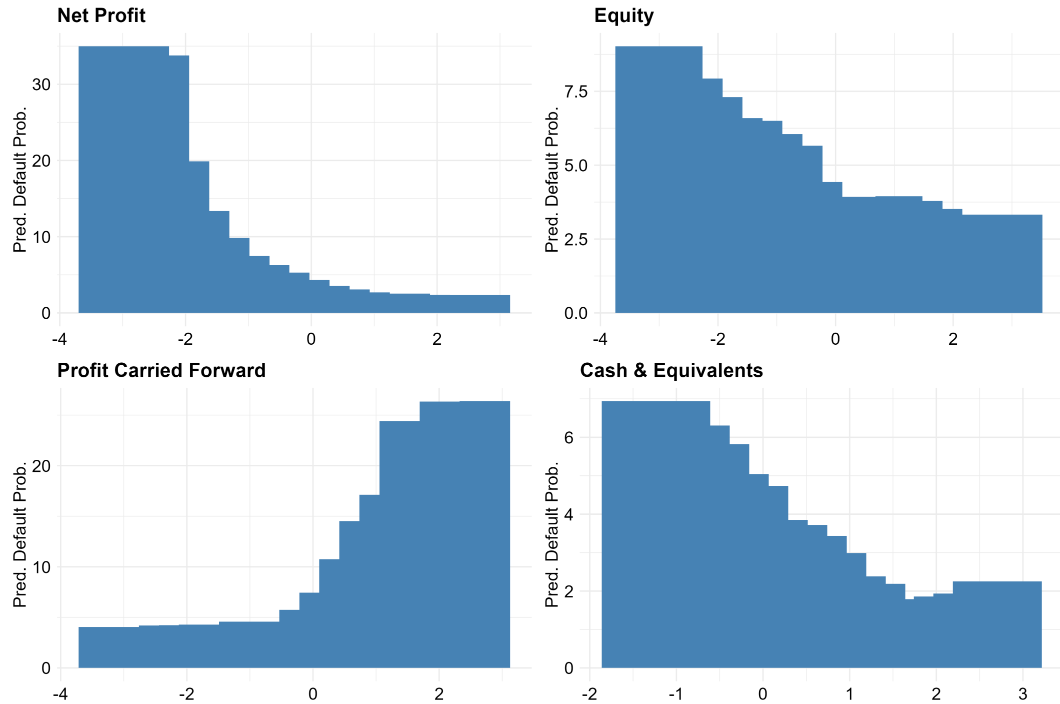

Abbildung 5.6: Marginal Response for the XGBoost Strategy A Model (without feature engineering) for the top 4 features.

Bivariate Response

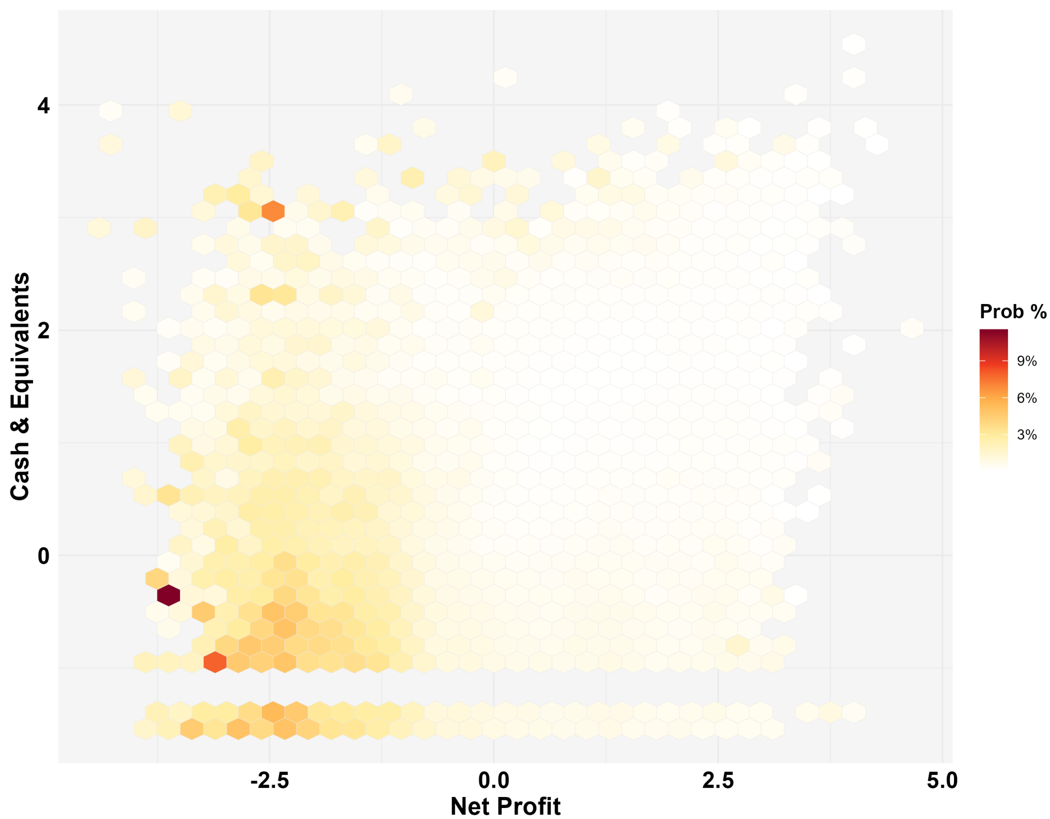

Abbildung 5.7: Bivariate Response for the XGBoost Strategy A Model (without feature engineering) for net profit and cash.

Error

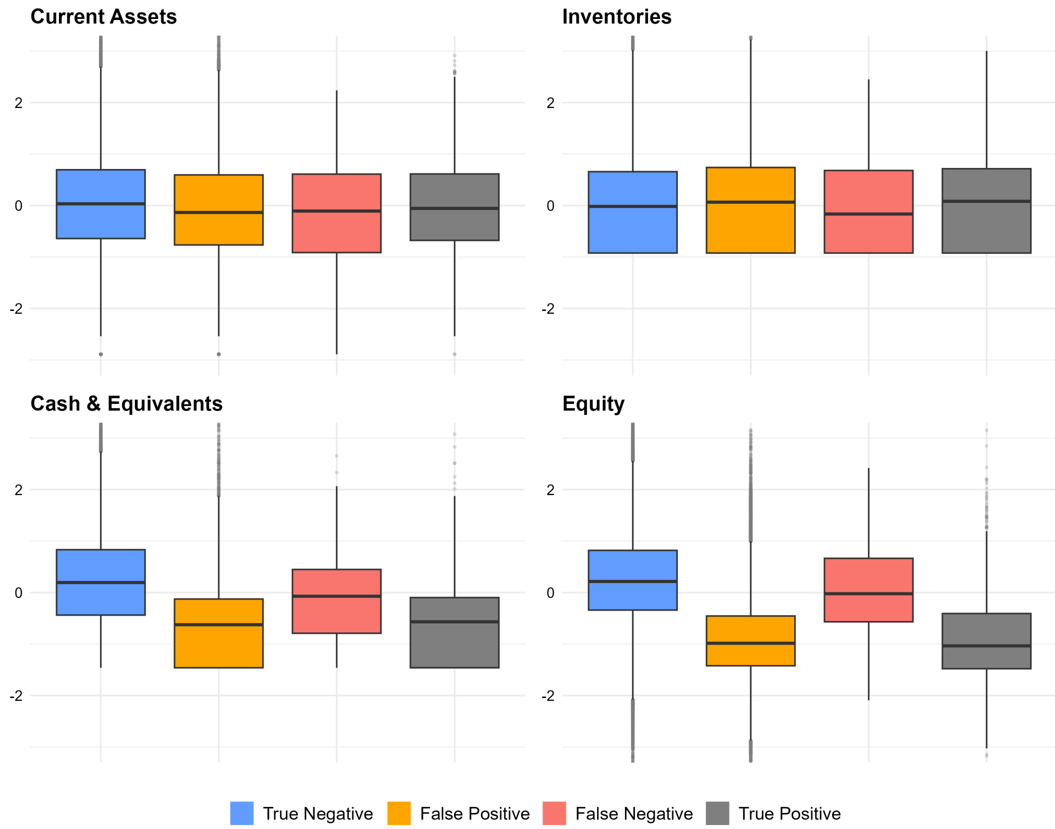

Abbildung 5.8: Error rate of the XGBoost Base Model (without feature engineering).

Tabelle 5.4: Forensic Feature Summary: Strategy A.

Error Type	Count	Net Profit	Equity	Profit C.F.	Cash & Equiv.	Curr. Assets
True Negative	132,930	0.2382	0.2154	0.2040	0.1946	0.0335
False Positive	36,258	-0.9745	-0.9848	-0.9039	-0.6246	-0.1346
False Negative	244	0.0322	-0.0224	0.0858	-0.0712	-0.1075
True Positive	1,222	-1.0563	-1.0352	-0.8372	-0.5679	-0.0556

Note: All financial feature columns reflect the median values for that group.

5.1.5 Model Insights: Strategy B

Error

Tabelle 5.5: Forensic Feature Summary: Strategy B

Error Type	Count	Net Profit	Equity	Cash & Equiv.	Profit C.F.	Curr. Assets
True Negative	127,957	0.2569	0.2338	0.2264	0.2231	0.0301
False Positive	41,231	-0.9109	-0.9078	-0.6246	-0.8132	-0.1107
False Negative	259	0.0032	-0.1304	-0.0784	0.0189	-0.2006
True Positive	1,207	-1.0751	-1.0524	-0.6246	-0.8480	-0.0384

Note: All financial feature columns reflect the median values for that group.

5.1.6 Model Insights: Strategy C

Error

Tabelle 5.6: Forensic Feature Summary: Strategy C

Error Type	Count	Net Profit	Equity	Cash & Equiv.	Profit C.F.	Robust Latent 5
True Negative	123,943	0.2894	0.2664	0.2497	0.2585	3.4105
False Positive	45,245	-0.8827	-0.8968	-0.6246	-0.8066	2.4265
False Negative	195	0.0644	-0.0286	0.1355	0.1407	3.2656
True Positive	1,271	-1.0154	-1.0174	-0.6246	-0.8066	2.4156

Note: All financial feature columns reflect the median values for that group.

5.1.7 Model Insights: Strategy D

Error

Tabelle 5.7: Forensic Feature Summary: Strategy D

Error Type	Count	Net Profit	Solvency Gap	Equity	Profit C.F.	Curr. Assets
True Negative	128,220	0.2594	-0.4610	0.2415	0.2231	0.0287
False Positive	40,968	-0.9109	0.7596	-0.9298	-0.8197	-0.1044
False Negative	231	0.0286	-0.2345	-0.0372	0.0341	-0.1277
True Positive	1,235	-1.0446	0.9123	-1.0352	-0.8197	-0.0523

Note: All financial feature columns reflect the median values for that group.

5.2 XGBoost Training Summary

The XGBoost training phase demonstrated the strong capability of gradient boosting to handle raw non-linearities, setting a high baseline for performance. Unlike the linear models, XGBoost naturally incorporated liquidity signals (Cash & Equivalents) alongside solvency metrics.

Table 5.8 summarizes the specific behavioral shifts introduced by each augmentation strategy.

5.2.1 Comparative Insights

The "Zombie Hunter" (Strategy D) Strategy D achieved the highest AUC (0.8138) not by catching the most defaulters, but by drastically reducing False Positives. The forensic analysis shows that Strategy D had $\sim$4,000 fewer False Positives than the Base Model (40,968 vs 44,580).

- **Why it works:** Gradient boosting trees approximate diagonal decision boundaries using step-functions. A feature like `Solvency Gap` ($Liabilities - Equity$) provides an explicit diagonal cut that is difficult for a shallow tree to learn from raw *Liabilities* and *Equity* features alone. This interaction term allowed the model to correctly classify "Zombie Companies"—firms with bad ratios that survive—thereby cleaning up the confusion matrix.

The Robustness of Noise (Strategy C) Among the deep learning strategies, Strategy C (Feature Denoising) was the most effective forensic tool. While Strategy A (Standard VAE) degraded performance by "smoothing" over liquidity risks (missing firms with negative cash), Strategy C's "Robust Latent 5" feature helped the model identify imposters.

- **Why it works:** By training on corrupted inputs, the Denoising Autoencoder learns to ignore transient fluctuations in accounting ratios. When fed into XGBoost, these robust features acted as a "validity check" on the raw financial data, preventing the model from overfitting to noise.

Tabelle 5.8: Summary of XGBoost Strategy Contributions and Failure Modes

Strategy	**Primary Risk Focus**	**False Negative Profile (Missed Risks)**	**Methodological Insight**
Base Model	**Liquidity** (Cash) & Profit	**The "Profit Illusion"**: Firms with Positive Profit and Positive Cash, but negative Equity.	The trees rely heavily on "Cash Burn." It misses firms that manipulate accounting profit to mask structural insolvency.
Strategy A(Dim. Reduction)	Smooth Manifold	**Liquidity Blindness**: Missed firms with *Negative Cash* (-0.071), which the Base Model caught.	**Over-Smoothing**: The VAE's KL-divergence constraint forces a smooth Gaussian latent space. This likely blurred the sharp, discrete "cliff" where low liquidity leads to default.
Strategy C (Denoising)	Robust Structure	**The "Deceptive Solid"**: Missed firms with High Profit (0.064) and High Cash (0.135).	**Noise Filtering**: The DAE successfully filtered out noisy financial ratios, allowing the model to focus on structural integrity. It only failed on firms that looked "perfect" on paper.
Strategy D (Manual Eng.)	**Solvency Interaction**	**Strategic Misses**: Accepted slightly more misses (FN) in exchange for massive FP reduction.	**The "Zombie" Filter**: The explicit "Solvency Gap" feature ($Liabilities - Equity$) allowed the model to slash False Positives by ∼4,000, distinguishing "distressed survivors" from actual defaulters.

The VAE Limitation (Strategy A) Strategy A underperformed the Base Model. The forensic data reveals that it missed 244 risks (compared to 175 for Base), specifically failing to detect firms with negative cash flow.

- **Why it failed:** The standard VAE optimizes for a continuous, smooth latent manifold (via the KL-Divergence loss). However, corporate default is often a discontinuous, "tail risk" event. By forcing the data into a smooth normal distribution, the VAE likely compressed the "dangerous outliers" into the "safe cluster," stripping away the sharp signal needed for crisis detection.

Kapitel 6

Training Results

6.1 Overall Insights: GLM vs. XGBoost

Comparing the forensic error profiles of the Generalized Linear Models (GLM) and the Gradient Boosting Machines (XGBoost) reveals a fundamental divergence in risk perception. While both models achieved comparable AUC scores (≈ 0.81), they arrived at these predictions through entirely different "financial philosophies."

6.1.1 Comparative Strengths and Weaknesses

GLM: The Solvency Architect The GLM excels at assessing **structural solvency**. By weighing *Total Assets* against *Liabilities*, it effectively identifies firms that are fundamentally underwater.

- **Where it struggles:** The GLM suffers from a "Linear Cancellation" effect. It systematically misses **Asset-Heavy / High-Debt** firms (False Negatives). Because the model is a linear equation ($w_1 \cdot Assets - w_2 \cdot Liabilities$), a massive pile of assets can mathematically "cancel out" massive debt, leading the model to classify a highly leveraged firm as safe. It lacks the ability to detect the "tipping point" where leverage becomes fatal regardless of asset book value.

XGBoost: The Liquidity Hawk XGBoost focuses intensely on **operational liquidity**. Its decision trees naturally prioritize "cliffs"—sharp thresholds in *Cash* and *Net Profit* where survival becomes impossible.

- **Where it struggles:** XGBoost is vulnerable to the "Profit Illusion." Its False Negatives are often firms with **Positive Accounting Profit** but negative equity. The trees see the positive profit and high cash flow and assume safety, failing to recognize that the firm is technically insolvent (Negative Equity). It misses the "Zombie" firms that GLM catches easily.

6.1.2 The Role of VAE and Deep Learning

The impact of the VAE architectures was diametrically opposed for the two model classes, highlighting the importance of alignment between feature engineering and model structure.

- **For GLM (High Impact):** The VAE (Strategy A) was highly beneficial. The GLM struggles with sparse, noisy ratios. The VAE compressed these into dense, correlated latent features, allowing the GLM to shift from a sparse Lasso selection to a dense Ridge penalization ($\alpha \approx 0.3$), effectively capturing the "shape" of the data that a simple linear line could not.
- **For XGBoost (Mixed Impact):** The standard VAE (Strategy A) was detrimental. The VAE's Gaussian constraint smoothed over the sharp "tail risks" that XGBoost needs to split on. However, the Denoising Autoencoder (Strategy C) was successful because it acted as a noise filter rather than a smoother, allowing XGBoost to ignore accounting noise and focus on robust structural signals.

6.1.3 Path Forward: The Ensemble Hypothesis

The most critical discovery is that **GLM and XGBoost are making different mistakes.**

- GLM misses the *High Asset/High Debt* firms.

- XGBoost misses the *Profitable/Insolvent* firms.

Since their error profiles are uncorrelated (orthogonal), they are ideal candidates for an **Ensemble Approach**. A simple mean average of their predicted default probabilities would likely "cancel out" these specific blind spots—the GLM would catch the insolvent zombies that fool XGBoost, and XGBoost would catch the liquidity crises that slip past the GLM's linear equation. This suggests that a hybrid "Expert Voting" system would yield higher robustness than any single model strategy alone.

6.2 Performance Leaderboard

Kapitel 7

Model Selection

7.1 Performance in the Test-set

7.1.1 GLM

Tabelle 7.1: GLM Model Performance Leaderboard (Test Set)

Model	**AUC**	**Brier Score (%)**	**Pen. Brier Score (%)**
Strategy A (Latent)	0.8171	0.8317%	1.257%
Strategy D (Residual Fit)	0.8163	0.8323%	1.258%
Base Model	0.8161	0.8322%	1.258%
Strategy B (Anomaly)	0.8160	0.8316%	1.257%
Strategy C (Regime)	0.8139	0.8320%	1.258%

7.1.2 Random Forest

7.1.3 Boosting

Tabelle 7.2: XGBoost Model Performance Leaderboard (Test Set)

Model	**AUC**	**Brier Score (%)**	**Pen. Brier Score (%)**
Strategy B (Anomaly Score)	0.8205	0.8288%	1.254%
Base Model	0.8203	0.8286%	1.254%
Strategy A (Dim. Reduction)	0.8192	0.8287%	1.254%
Strategy D (Manual Feature Eng.)	0.8185	0.8279%	1.254%
Strategy C (Feature Denoising)	0.8034	0.8309%	1.257%

7.2 Ensemble Approach

Tabelle 7.3: Final Ensemble Performance Leaderboard (Test Set)

Model Strategy	**AUC**	**Brier Score (%)**	**Pen. Brier Score (%)**
Ensemble C (Selected Top)	**0.8268**	**0.8278%**	**1.253%**
Ensemble B (All Model Avg)	0.8265	0.8287%	1.254%
Ensemble A (Base Avg)	0.8265	0.8281%	1.254%
Single Best (XGB Strat B)	*0.8205*	*0.8288%*	*1.254%*

7.3 Model Selection & Summary