ILAB: OeNB

Documentation

Variational Autoencoders

Author:
Tristan Leiter

12. Februar 2026

Inhaltsverzeichnis

Kapitel 1

Variational Autoencoder (VAE) Integration & Diagnostics

1.1 Motivation and Objective

The primary objective of integrating a Variational Autoencoder (VAE) into the credit risk modeling pipeline was to leverage unsupervised learning to capture non-linear latent structures that standard gradient boosting models (XGBoost) might miss. While XGBoost excels at identifying decision boundaries based on provided features, it struggles to capture the underlying "manifold" of the data without explicit feature engineering.

The implementation follows a "Hybrid Expert" approach, where the VAE acts as a feature extractor and anomaly detector to augment the supervised XGBoost model. The specific goals were:

- **Dimensionality Reduction:** Compressing 11 financial ratios into a lower-dimensional latent space to identify clusters of firm behavior.
- **Anomaly Detection:** Using reconstruction error to flag firms that are structurally "weird" or inconsistent with the general population, hypothesizing that high structural anomaly correlates with default risk.
- **Residual Modeling:** Using the VAE to specifically model the "hard samples" that the baseline XGBoost model fails to predict correctly.

1.2 VAE Implementation Strategies

Several strategies were tested to extract value from the VAE. The code iterates through five distinct approaches (Strategies A through E) to determine the optimal integration method.

1.2.1 Strategy A: Latent Features (Dimensionality Reduction)

This strategy utilizes the **Encoder** part of the VAE. The 11 continuous financial variables and categorical metadata are compressed into an 8-dimensional latent vector ($l_1 \dots l_8$).

- **Mechanism:** The encoder maps the input x to a latent distribution $P(z|x)$. The mean of this distribution is extracted as the new feature set.
- **Hypothesis:** These latent features capture non-linear "concepts" (e.g., a "high-growth but low-liquidity" cluster) that are more robust than raw financial ratios.

1.2.2 Strategy B: Anomaly Score (Reconstruction Error)

This strategy utilizes the **Decoder**. It measures how well the VAE can reconstruct the input data after compression.

- **Mechanism:** The model calculates the Mean Squared Error (MSE) for continuous variables and Binary Cross Entropy (BCE) for categorical variables between the input and the reconstruction. The sum is the `anomaly_score_total`.
- **Refinement (Strategy B Revised):** The anomaly score was augmented with manual "Zombie" interaction features, such as the `Gap_Debt_Equity` and `Ratio_Cash_Profit`, to capture specific diagonal decision boundaries that pure reconstruction error might miss.

1.2.3 Strategy C: Regime Switching (Directional Surprise)

Instead of a single scalar error score, this strategy looks at the *direction* of the error.

- **Mechanism:** Signed residuals are calculated ($Actual - Reconstructed$). This creates "Soft Surprise" features.
- **Hypothesis:** If a firm has much higher profit ($f8$) than the VAE expects based on its other features, this "positive surprise" contains signal distinct from the raw profit value.

1.2.4 Strategy D: Residual Fit (The Specialist VAE)

This is a "Hard Sample Mining" approach.

- **Mechanism:** A base XGBoost model is trained, and the residuals (absolute errors) are calculated. A "Specialist VAE" is then trained *only* on the top 25% "hardest" samples (the observations with the highest residuals).
- **Usage:** This Specialist VAE scores the full dataset. A low reconstruction error from this model indicates the firm looks like a "hard failure case," providing a `Specialist_Risk_Score`.

1.2.5 Strategy E: Feature Denoising (DAE)

This strategy employs a Denoising Autoencoder (DAE) to force robustness.

- **Mechanism:** A dropout layer (rate 0.1) is added to the input of the encoder. The model attempts to reconstruct the clean original data from the corrupted input.
- **Output:** "Robust" latent features (`dae_l1` ... `dae_l8`) are extracted, theoretically effectively filtering out noise from the financial ratios.

1.3 Residual Diagnostics Methodology

A critical finding during the analysis was the failure of the standard decision threshold (0.5) due to the extreme class imbalance (approx. 0.86% default rate). A new diagnostic methodology was developed to evaluate model performance forensically.

1.3.1 Dynamic Thresholding (Youden's J)

The standard classification threshold resulted in near-zero True Positives. The new methodology replaces the arbitrary 0.5 cutoff with an optimized threshold derived from the ROC curve.

- **Procedure:** The code calculates the Receiver Operating Characteristic (ROC) curve and extracts the coordinates for the "best" threshold (Youden's J statistic), which maximizes the sum of Sensitivity and Specificity.
- **Result:** This shifts the decision boundary (e.g., to 0.02), allowing the model to capture the majority of defaults at the cost of higher false positives, rendering the model usable for risk screening.

1.3.2 Forensic Error Analysis

To understand *why* the model fails, the predictions are segmented into four confusion matrix categories (True Positive, True Negative, False Positive, False Negative).

- **Stealth Defaulters (False Negatives):** The analysis compares the distribution of key drivers (Profit, Solvency Gap) for False Negatives against True Positives. This reveals risks that are invisible to the model (e.g., profitable firms that default due to liquidity shocks).
- **Healthy Losers (False Positives):** The analysis compares False Positives against True Negatives. This identifies "Zombie" firms—those with high debt and low profit that technically *should* default but survive due to hidden buffers like high cash reserves.

Kapitel 2

Final Model Optimization: The Solvency Stabilizer

2.1 The Breakthrough: Solving the "Healthy Loser" Panic

The primary weakness identified in previous iterations was the model's tendency to panic when observing negative profitability, resulting in a high rate of False Positives. This phenomenon, termed the "Healthy Loser" Panic, flagged firms that were technically unprofitable but structurally sound due to high cash reserves.

By implementing **Strategy D**, which utilizes the `Feature_Stabilizer` (switching the focus to Cash $f5$ when Profit $f8$ is negative), we achieved a "Double Win": the model successfully caught more defaulters while drastically reducing false alarms.

2.2 Quantitative Impact: Before vs. After

The implementation of the stabilizer logic produced a decisive shift in model performance. Table 2.1 details the impact on the confusion matrix.

Tabelle 2.1: Impact of Strategy D (Feature Stabilizer) on Model Performance

Metric	Previous Result	Strategy D Result	Improvement
True Positives (Caught Risks)	1,191	**1,361**	+170 (Sensitivity ↑)
False Negatives (Missed Risks)	275	**105**	-170 (Safety ↑)
False Positives (False Alarms)	39,738	**26,077**	-13,661 (Precision ↑)
True Negatives (Correct Safe)	129,450	**143,111**	+13,661 (Efficiency ↑)

Key Outcomes:

- **Efficiency:** 13,661 healthy firms were rescued from being wrongly flagged as risky.
- **Sensitivity:** Simultaneously, an additional 170 defaulters were caught that were previously missed.
- **Total Capture Rate:** The model now captures **92.8%** of all defaults ($1,361/1,466$), a figure considered exceptional for credit risk modeling.

2.3 Mechanism of Action: Why the Stabilizer Worked

The significant drop in False Positives (from 39k to 26k) confirms the hypothesis regarding "Healthy Losers."

- **The Old Logic:** The model observed $f8$ (Profit) at values like -2.0 and immediately classified the firm as high risk based on income statement bleeding.
- **The New Logic:** The `Feature_Stabilizer` intervened: *"Profit is negative; therefore, evaluate Cash ($f5$) instead."*

- **The Result:** For the 13,661 rescued firms, the model recognized the "Cash Cushion" ($f5 > 0$) as a sufficient buffer against the "Profit Bleed," correctly reclassifying them as Survivors (True Negatives).

2.4 Forensic Analysis of Remaining Errors

Despite optimization, specific error groups remain. A forensic analysis reveals the distinct profiles of these residual errors.

2.4.1 The Remaining False Negatives (105 Firms)

These are the **"Stealth Defaulters."**

- **Profile:** Median Profit ($f8$) ≈ -0.157; Median Solvency Gap ≈ -0.235.
- **Diagnosis:** These firms exhibit "safe" balance sheets (Equity > Debt) and are barely losing money.
- **Root Cause:** Financially, they appear identical to safe firms. Their default is likely driven by non-financial factors—such as fraud, lawsuits, or sudden management exits—or extreme short-term liquidity shocks not captured in annual reports. This represents the likely "irreducible error" floor for a model based solely on annual financial statements.

2.4.2 The Remaining False Positives (26,077 Firms)

These are the **"Hardcore Zombies."**

- **Profile:** Massive Losses (Median $f8 \approx -1.05$); Massive Debt (Median Gap $\approx +0.90$).
- **Diagnosis:** By all standard financial logic, these firms should be insolvent.
- **Root Cause:** Their survival is likely due to external factors such as government bailouts, parent company guarantees, or extreme asset liquidation.
- **Interpretation:** Flagging these firms is the *correct behavior* for a risk model. They represent high risk; they simply survived against the odds.

2.5 Future Improvements and Limitations

While the current model is production-ready, further reduction of the remaining error types requires data beyond the current scope.

2.5.1 Addressing False Positives (The "Zombie" Survivor)

The model fails to exclude these firms because their survival is mathematically improbable based on their financials alone.

- **Limitation:** The model cannot see "external support."
- **Solution path:** To reduce these False Positives, we would need to integrate:
  1. **Ownership Structure Data:** Identifying parent companies with deep pockets.
  2. **State Aid/Subsidy Data:** Identifying firms receiving government support.
  3. **News Sentiment:** Detecting announcements of restructuring or bailouts.
- **Recommendation:** Treat these 26,077 firms as a "High Risk Watchlist." They are not defaults yet, but they are living on borrowed time.

2.5.2 Addressing False Negatives (The "Stealth" Defaulter)

The model fails to catch these firms because their annual reports look healthy right up until the moment of collapse.

- **Limitation:** The latency of annual reporting masks sudden liquidity crises or fraud.
- **Solution path:** To capture these Stealth Defaulters, we would need:
    1. **Higher Frequency Data:** Monthly cash flow or bank transaction data to catch sudden liquidity drying.
    2. **Fraud Detection Models:** Applying Benford's Law or forensic accounting ratios to detect manipulated financial statements.
    3. **Legal Filings:** Monitoring court dockets for sudden litigation which often precedes "healthy" defaults.

2.6 Final Recommendation

We have reached the point of diminishing returns for feature engineering on this specific dataset. The recommendation is to **Stop Engineering** and accept the model. Capturing ∼93% of defaults with a drastically reduced False Positive rate is a robust result. The pipeline (Strategy D + Threshold Optimization) should now be applied to the final Holdout Test Set to confirm these metrics on unseen data.