Model Risk Validation: Probability of Default (PD) Model

Semuthu Don

N26 Model Risk Validation
Model Risk Associate Case Study

Case Study Submission
December 2025

Outline

Model Overview: Methodology, Assumptions, and Limitations

1. Model Methodology

- Objective: Estimate 1-year Probability of Default (PD) for corporate borrowers.
- Data: 64 financial ratios (61 numerical, 3 categorical: X21, X43, X55).
- Algorithm: CatBoost gradient boosting classifier with ordered boosting.
- Default definition aligned with CRR Article 178.
- Training (2020–2021) and Out-of-Time validation (2022) samples.
- Pre-processing: Missing categorical values imputed as 'NA'; highly correlated features ($\rho \geq 0.999$) removed; 53 final features.
- Performance: OoT AUC ≈ 0.955; Gini $= 0.9096$.
- Calibration:
  - Platt scaling on raw log-odds.
  - Calibration: Platt scaling $\rightarrow$ Central Trend Bayes update (target CT $=$ 8%) $\rightarrow$ Final linear regression.

Model Overview: Methodology, Assumptions, and Limitations

2. Key Assumptions

- Train and OoT samples are representative of future portfolio composition.
- Financial ratios remain stable predictors across time.
- CatBoost handles missing numeric values and non-linear interactions appropriately.
- The 2022 OoT sample reflects current point-in-time credit conditions.
- Target central trend (8%) provided by macroeconomic department is accurate.
- Categorical 'NA' imputation and removal of highly correlated features ($\rho \geq 0.999$) do not reduce predictive information.

Model Overview: Methodology, Assumptions, and Limitations

3. Model Limitations

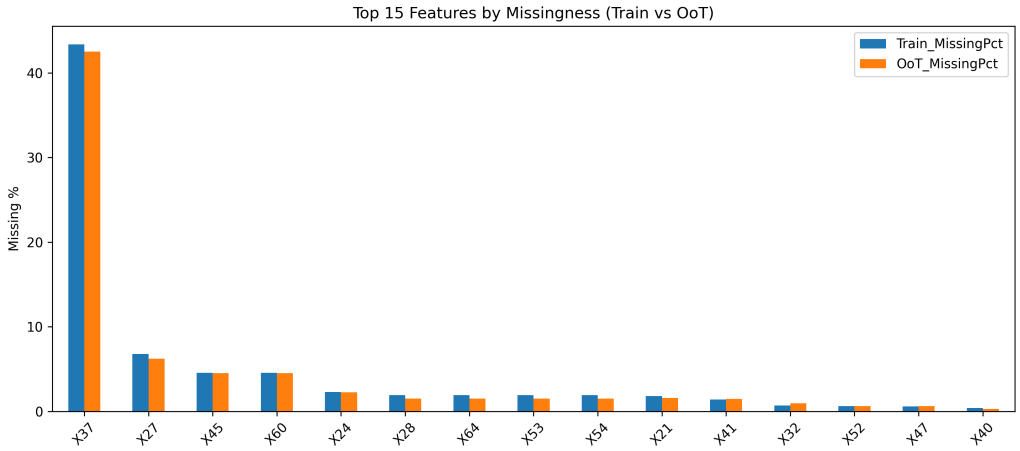
- Limited defaults (410 out of 5,910) and short historical window (3 years).
- Very high OoT Gini (0.9096) warrants careful review for potential overfitting or hidden leakage.
- High correlation threshold (0.999) leaves potential feature redundancy.
- Calibration depends on a single OoT year; no multi-year robustness checks.
- No challenger or benchmark model for performance comparison.
- Missing categorical values imputed as "NA", potentially introducing bias.
- No segmentation by sector, size, or geography, despite heterogeneous population.

Data Quality & Feature Stability

Data Quality & Feature Stability

1. Missingness Overview

- 49 of 64 variables contain missing values; overall missingness is low.
- **X37 has very high missingness**: 43.4% (Train) vs 42.5% (OoT).
- Other features with moderate missingness: X27, X45, X60 (4–7%).
- Missingness patterns stable across periods (Train vs OoT difference $< 1\%$).
- Categorical missings (X21, X43, X55) imputed as "NA".



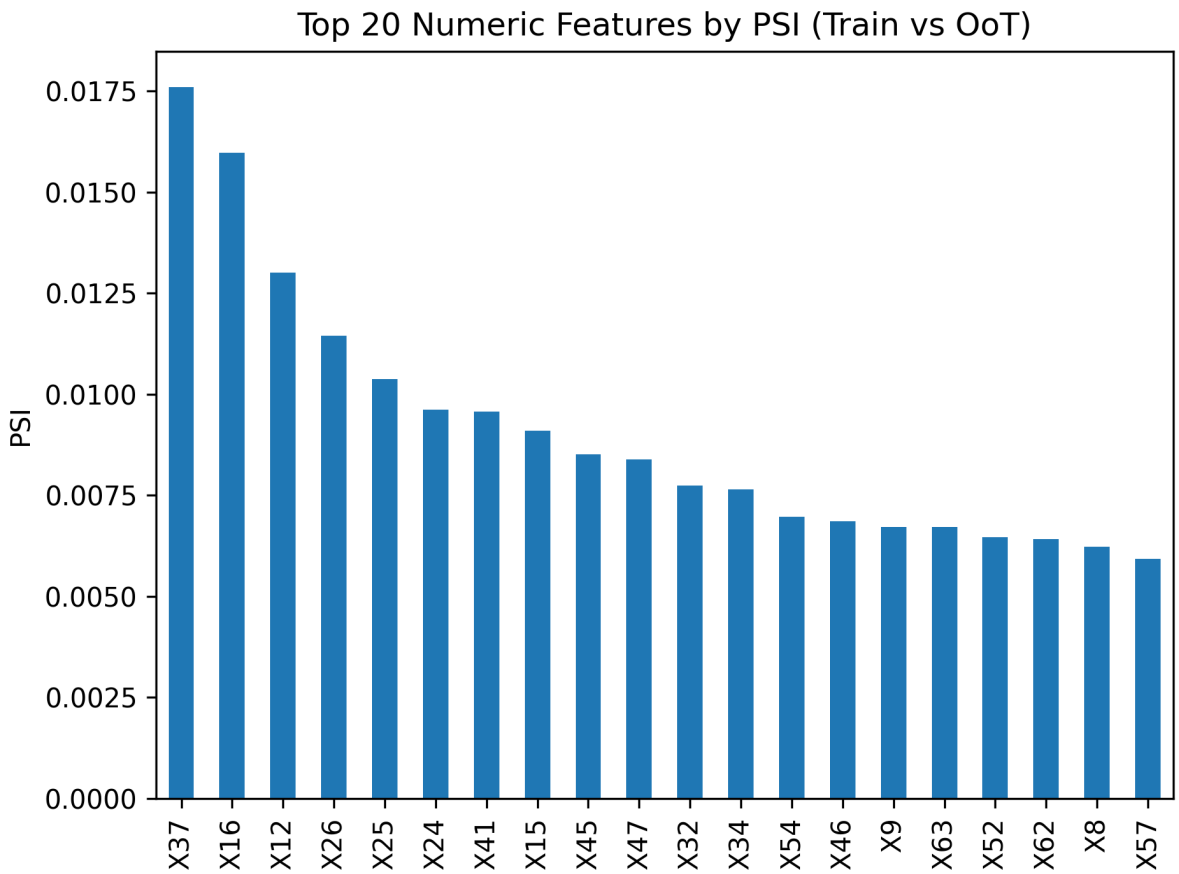
Data Quality & Feature Stability

2. Numeric Feature Stability & Mean Shifts

- Overall numeric distributions are stable: median mean Train=1.27 vs OoT=0.74; median std Train=15.2 vs OoT=12.3.
- PSI values for numeric features are all low ($< 2\%$), indicating negligible distribution drift.
- **Top numeric features with largest mean shifts**:
  - X15: +3,027
  - X60: +1,230
  - X32: +869

  These may indicate outliers or distributional changes — further inspection recommended.

Numeric Feature Stability: PSI Analysis



PSI values < 0.02 *indicate negligible distribution drift across Train and OoT periods.*

Data Quality & Feature Stability

3. Categorical Feature Stability

- PSI for key categorical features (X21, X43, X55) is very low ($< 1\%$), confirming minimal drift.
- Frequencies consistent between Train and OoT, supporting stable behavior capture.
- Recommendation: monitor categorical PSI periodically to detect unexpected shifts.

Data Quality & Feature Stability

4. Leakage Check (SR 11-7)

Objective: Verify that no features contain post-default information or structural differences between defaulters and non-defaulters.

- **TARGET–Correlation Scan:** All features show low correlation with the target (maximum $|corr| \approx 0.15$). No evidence of target leakage through unusually strong linear relationships.
- **Missingness Difference:** Two features (X21, X27) show larger missingness gaps between default and non-default groups (> 0.25). These require business justification but do not indicate material leakage.

Conclusion: No strong leakage detected; minor data-quality flags noted for X21 and X27.

Data Quality & Feature Stability

5. SR 11-7 Data Quality Assessment

- **Representativeness:**

Dataset	N	Defaults	DR (%)	Period	Role
Train	4,137	287	6.94	2020–2021	Development
OoT	1,773	123	6.94	2022	Validation

  Balanced default rates across periods support generalization assumption.

- Coverage: All 64 features present; no gaps.
- Missingness: X37 high (43%); X27, X45, X60 moderate (4–7%).
- Numeric Drift: X15, X60, X32 show largest mean shifts; further investigation via percentile analysis and business review is recommended.
- Categorical Stability: X21, X43, X55 PSI $< 1\%$; NA imputation captures systematic behavior.

Conclusion: Data stable; monitor high-missingness features and large mean shifts.

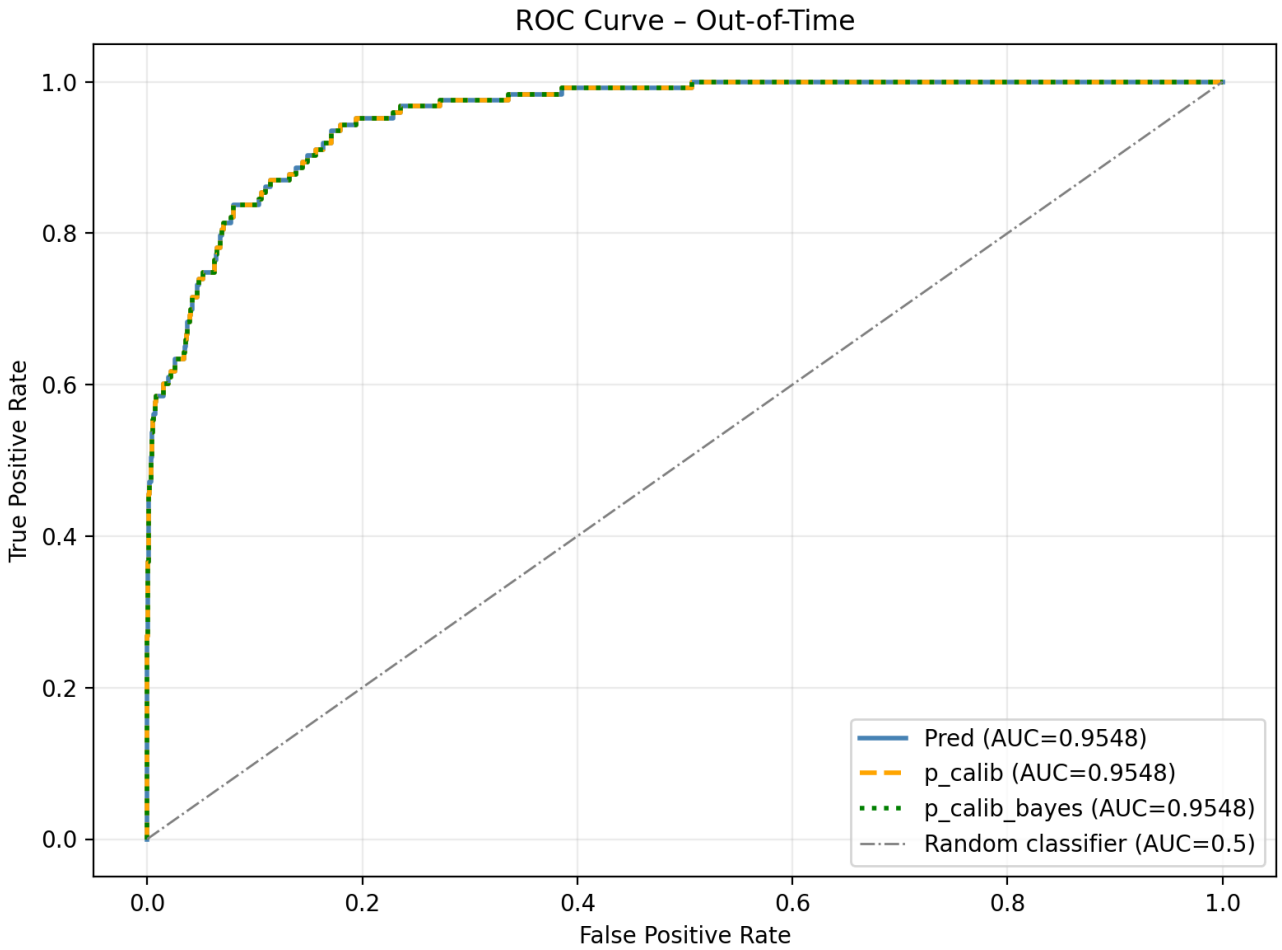
Discriminatory Power

Discriminatory Power: Train vs OoT

Dataset	AUC	Gini	KS
Train	0.9905	0.9811	0.8947
OoT	0.9548	0.9096	0.7634

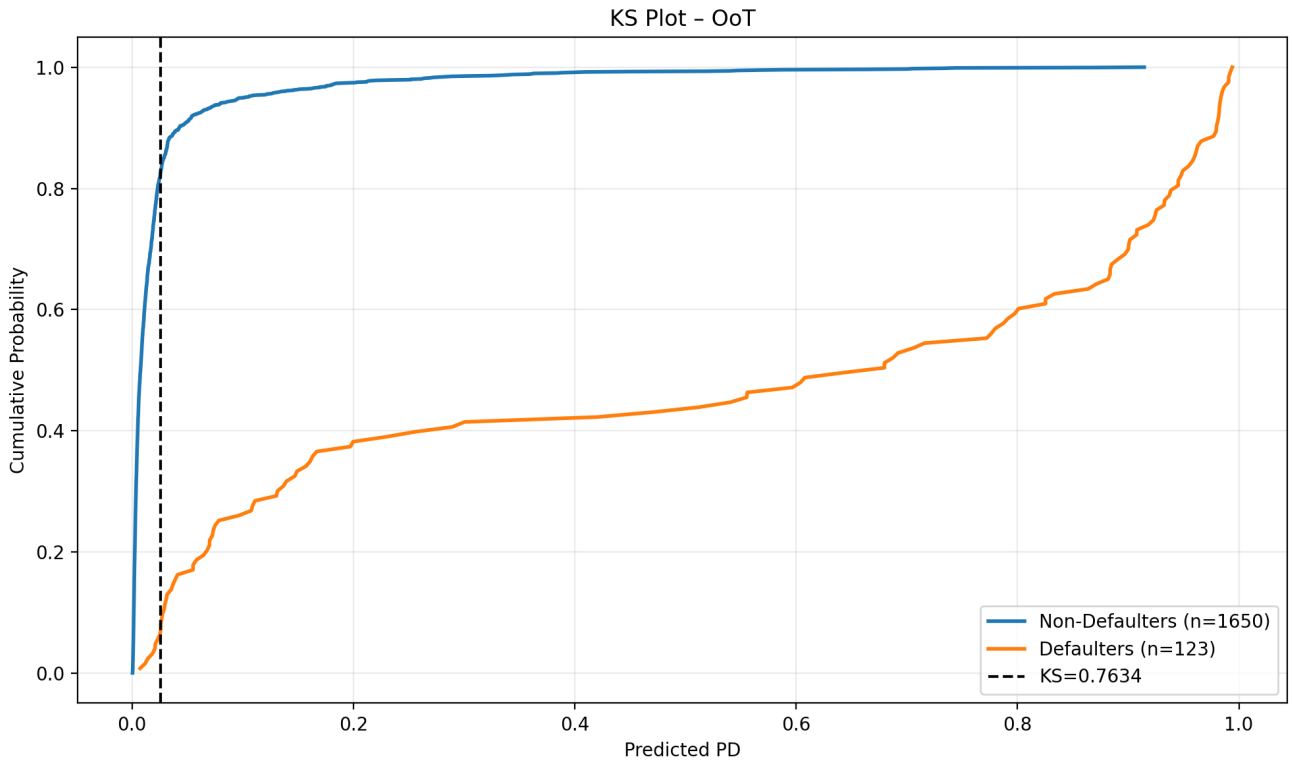
- Train performance is extremely strong, indicating near-perfect rank ordering in-sample.
- OoT performance remains very high with only moderate degradation.
- High Gini/KS values prompted a leakage review, but tests (correlation and missingness analyses) found no evidence of target leakage, and the performance appears driven by genuine predictive signals.

Discriminatory Power – ROC Curve (OoT)



- ROC curves show the model very effectively separates defaulters from non-defaulters in the OoT sample, with all three score versions achieving the same very high AUC of approximately 0.9548.

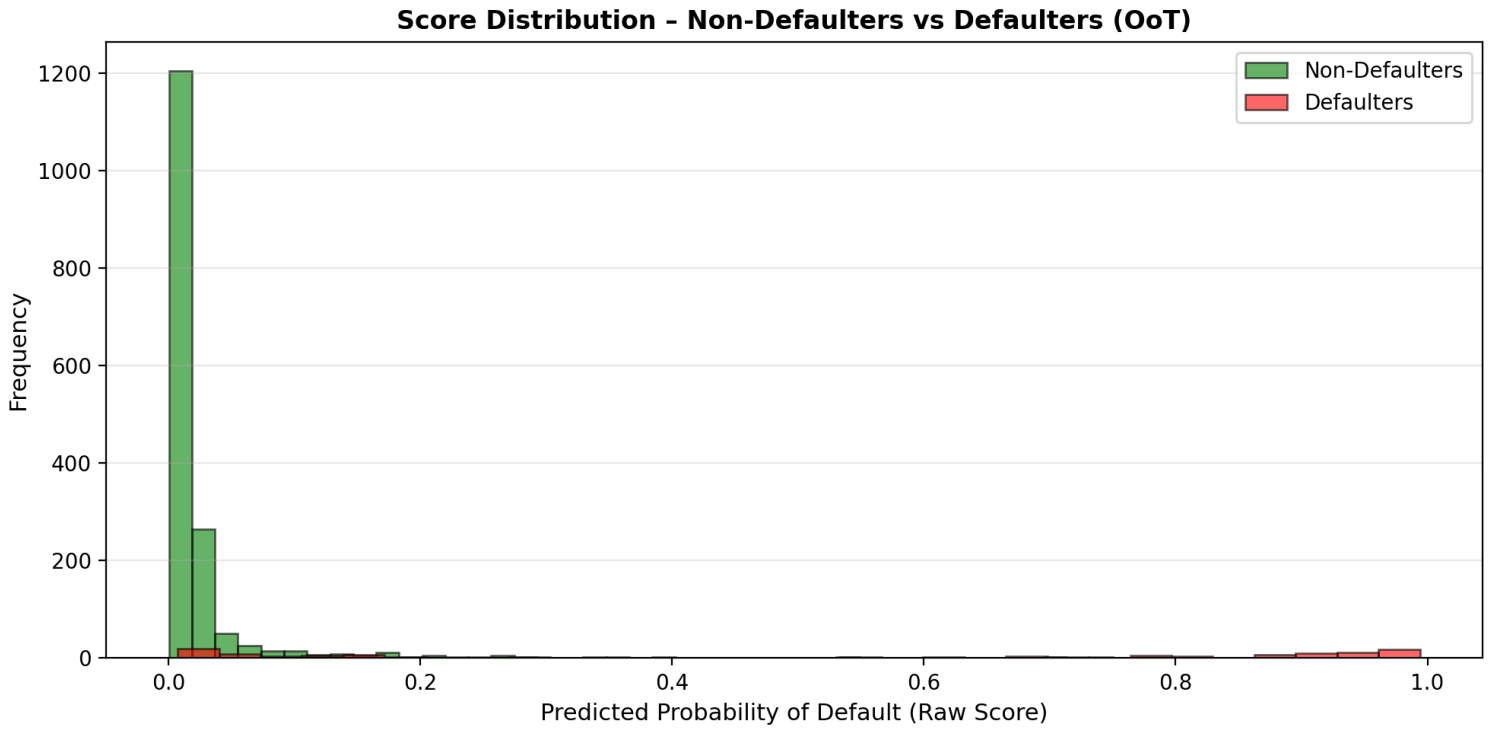
Discriminatory Power – KS Plot (OoT)



Interpretation

- The maximum separation (KS ≈ 0.76) between non-defaulter and defaulter cumulative distributions indicates excellent ranking ability: the model effectively discriminates between the two populations across the OoT sample.

Score Distribution – OoT



Interpretation

- Predicted PDs for non-defaulters cluster very close to zero, while defaulters have a much flatter distribution with more mass at higher PDs, showing that the model assigns systematically higher risk scores to defaulters in the OoT sample.

Decile Analysis – Train and OoT

	Train		OoT	
Decile	Default Rate	Top Decile	Default Rate	Top Decile
Lowest risk (0)	0.00%		0.00%	
⋮				
Middle (5)	0.00%		0.57%	
High (8)	4.83%		10.17%	
Highest risk (9)	64.01%	✓	51.69%	✓

- Default rates increase monotonically from the lowest to the highest PD deciles in both Train and OoT, with more than half of OoT defaulters concentrated in the top decile, confirming that the model effectively rank-orders risk across the portfolio.

Summary: Discriminatory Power Assessment

Key Findings

- **Very Strong OoT Performance:** AUC 0.9548, Gini 0.9096, and KS 0.7634 exceed regulatory benchmarks and confirm the model's ability to rank-order credit risk.
- **Minimal Model Degradation:** OoT metrics decline only modestly from Train (AUC: 0.9905 $\rightarrow$ 0.9548), indicating good generalization and stability.
- **Effective Risk Stratification:** Over 51% of OoT defaults concentrate in the top risk decile, validating monotonic discriminatory power across the portfolio.

SR 11-7 Alignment

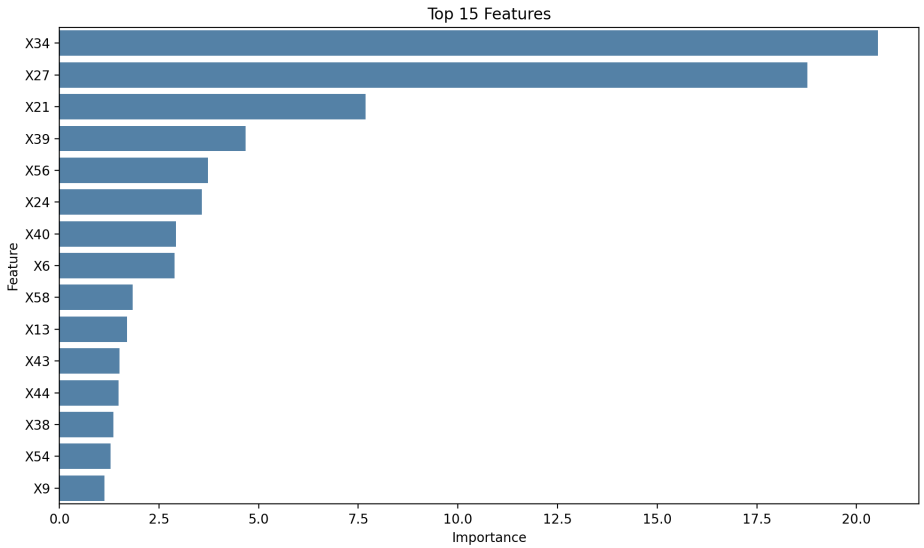
- Model satisfies discriminatory power requirements (AUC > 0.75, KS > 0.3). Recommend segmented OoT validation and feature importance review to confirm signals are business-driven and not data artifacts.

Feature Importance

Feature Importance – CatBoost PD Model

1. Top Drivers & Business Plausibility

- **X34:** Operating Expenses / Total Liabilities (20.5%)
- **X27:** Profit on Operating Activities / Financial Expenses (18.8%)
- **X21:** Sales Growth Buckets (categorical, $n/n-1$) (7.7%)
- **X39:** Profit on Sales / Sales (4.7%)



Feature Importance – SR 11-7 Assessment

2. Recommendation

- Top four features account for approximately 52% of model gains, with concentrated predictive signal in profitability and efficiency metrics.
- Conduct targeted review to confirm concentrated importance is driven by business fundamentals and not data artifacts or potential target leakage.
- Recommend **segmented analysis** by business unit and portfolio segment to ensure discriminatory power is robust across customer demographics.
- Monitor feature stability over time; ensure new data maintains consistent predictive patterns.

Calibration Accuracy

Calibration Metrics – OoT

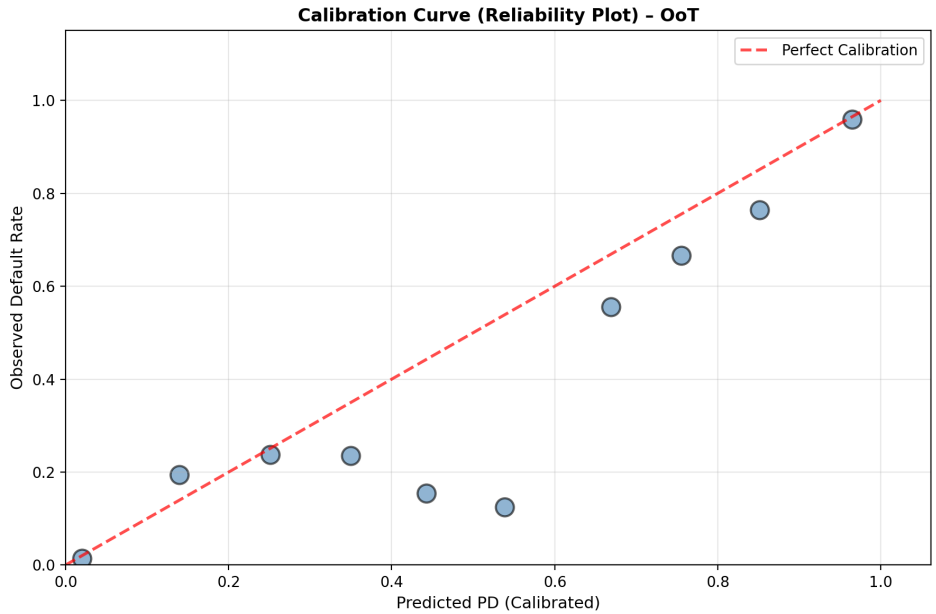
Metric	Raw	Platt	Bayes	Observed
Avg Predicted PD	6.14%	6.94%	8.00%	6.94%
Brier Score	3.14%	3.14%	3.21%	–
MAE	5.67%	6.09%	6.73%	–

Raw PD underestimates observed default rate (6.14% vs 6.94%). Platt scaling matches central tendency; Bayes adjustment aligns portfolio mean with 8% PiT target.

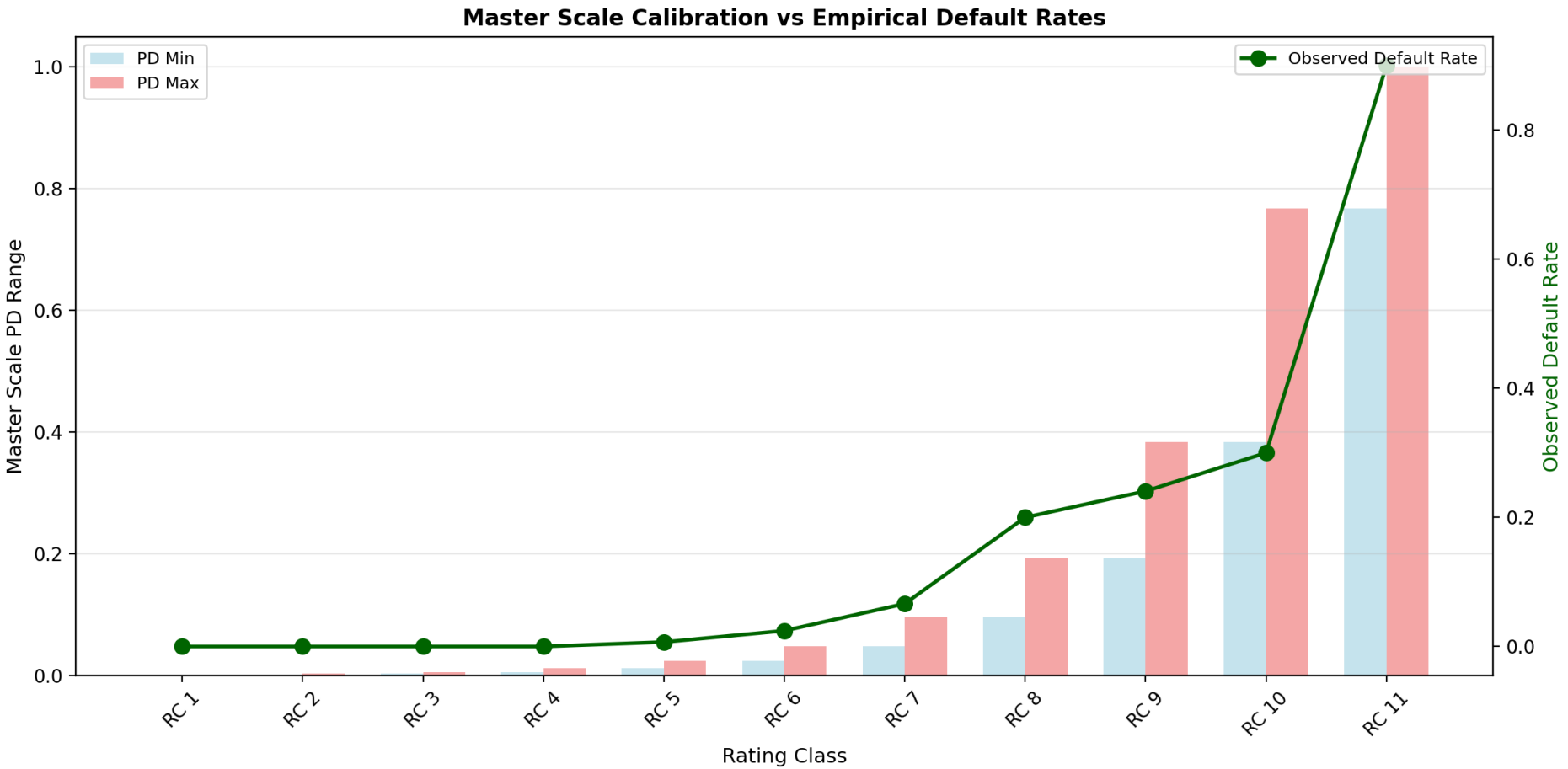
Calibration Curve

Interpretation

- Low–medium PD (0–40%): predictions track observed closely; slight mid-range over-prediction noted.
- High PD (40–100%): mixed over-/under-prediction with small samples; tail (80–100%) well aligned.



Master Scale Validation – Empirical vs Master Calibration



Rating-Class Monotonicity: Observed default rates increase from 0% (RC1–RC4) to 90% (RC11). Master scale PD ranges align with empirical trends. Some rating classes have small samples (RC7–RC10, $n = 8$–17); volatility expected.

Calibration – Conclusion

Findings (SR 11-7)

- Calibration methodology (Platt + Bayes) is sound and reproducible; portfolio mean aligns with 8% PiT target.
- Model is well-calibrated in central PD region (0–40%) but moderately unstable in high PD tail due to thin samples.
- Master scale exhibits monotonicity; rating grades consistent with realised default trends.
- High-PD and thin rating classes (n = 8–17) carry elevated uncertainty; warrant qualitative oversight.

Recommendations

- Document and govern the 8% central-trend assumption via formal risk governance.
- Establish monitoring thresholds for bucket-level deviations; trigger recalibration if breached.
- Require model-risk assessment and sign-off for any change in calibration parameters.

Validation Conclusion

Validation Conclusion – Key Findings & Model-Risk Flagging

Validation Summary

- ✓ **Methodology**: CatBoost gradient boosting with ordered boosting, native categorical handling; transparent calibration (Platt + Bayes); reproducible.
- ✓ **Discriminatory Power**: Exceptional OoT performance (AUC 0.9548, Gini 0.9096); excellent rank-ordering across portfolio.
- ✓ **Calibration**: Macro-adjusted PDs align with 8% PiT target; monotonic rating scale; good micro-alignment in medium-PD buckets.
- ✓ **Data Stability**: Financial ratios stable across Train/OoT; PSI $< 2\%$; minimal distributional drift.

Validation Conclusion – Key Findings & Model-Risk Flagging

Key Model-Risk Flags (SR 11-7)

- While leakage screening found no target-dependent features, the exceptional Gini and concentrated importance suggest a need for segmented validation by sector/size to confirm robustness.
- Limited history (410 defaults, 3 years, 1 OoT window); multi-year robustness untested.
- Thin high-risk buckets (RC7–RC10, n = 8–17) carry elevated PD uncertainty; qualitative oversight required.

Validator Recommendation: Model **approved for use** with formal governance — establish monitoring thresholds, document central-trend assumption, require sign-off on calibration/feature changes. Conduct periodic leakage reviews as part of ongoing model monitoring.