Micro-Level Indicators and Machine Learning as an Early Warning Signal Tool: Predicting Loan Defaults with Support Vector Model vs Logistic Regression

Prepared by: Jessica P. Hutalla

Abstract

The global financial system has undergone significant transformation, introducing new sources of risk that heighten the importance of timely credit risk monitoring for financial stability. Traditional macroprudential surveillance frameworks rely primarily on aggregate indicators and periodic stress tests, which may fail to capture borrower-level vulnerabilities in real time. This study emphasizes the role of micro-level indicators derived from granular loan-level data—including credit scores, repayment histories, collateral values, and loan characteristics—in predicting loan defaults and enhancing early-warning capabilities.

A comparative analysis is conducted between two supervised machine learning classification models: Logistic Regression and Support Vector Machines (SVM). Preliminary results reveal substantial differences in predictive performance. Logistic Regression demonstrates reasonable accuracy for the majority class (precision 0.79, recall 0.85, F1-score 0.82) but performs poorly in identifying default cases (precision 0.49, recall 0.39, F1-score 0.43). In contrast, the SVM achieves near-perfect performance across both classes, with precision, recall, and F1-scores exceeding 0.96, indicating strong discriminatory power and balanced classification.

These findings suggest that while Logistic Regression serves as a useful baseline model, SVMs offer significantly stronger predictive capability for loan default detection. The study highlights the potential of micro-level predictive models to support macroprudential early-warning frameworks and provides a foundation for future research on integrating borrower-level risk measures into systemic risk monitoring tools.

Table of Contents

I. Introduction

Central banks play a pivotal role in safeguarding financial stability through the continuous monitoring of key global and domestic risks and by maintaining the capacity for timely intervention, both in periods of stress and during normal economic conditions. Effective surveillance is essential to prevent localized financial vulnerabilities from propagating into system-wide crises. Among the risks under scrutiny are credit-related risks, particularly loan defaults. Together with counterparty risk, these vulnerabilities are central concerns in credit markets, as they significantly contribute to the amplification of financial shocks through balance sheet deterioration, ultimately giving rise to liquidity constraints and broader systemic instability.

While macroprudential surveillance has traditionally relied on aggregate indicators—such as stress tests based on system-wide banking balance sheet measures—these approaches may lack sensitivity to emerging micro-level risks that often precede broader financial instability. As a result, there is increasing interest in integrating predictive analytics into supervisory toolkits, with the aim of enhancing early-warning

capabilities during normal economic conditions and improving preparedness for potential crisis episodes.

This study highlights the importance of micro-level indicators in the assessment of credit risk and introduces a comparative analysis of two machine learning models: the Support Vector Machine (SVM) and the Logistic Regression model. Both are supervised learning algorithms widely used for classification tasks. Using granular loan-level data—including borrower credit scores, repayment histories, collateral values, and loan-specific characteristics—these models are employed to estimate borrower-level probabilities of default, providing a more detailed and forward-looking perspective on credit risk[1].

The best model or estimator in terms of accuracy and precision will be subject for deployment as a surveillance tool at the bank and system levels. The outputs are then aggregated into risk indicators that can inform macroprudential decision-making, such as inputs to the Systemic Vulnerability Index[2]. Specifically, the model aims to anticipate shifts in the risk distribution—such as increases in the tail of predicted default probabilities or concentration of risk in specific portfolios—that may signal fragility and warrant closer supervisory attention. The approach complements top-down stress testing by providing bottom-up signals rooted in borrower behavior and loan underwriting characteristics.

The contribution of this research idea is supposedly twofold. First, it **compares the construction and validation of an SVM-based vs the traditional logistic regression** default prediction model tailored for ongoing surveillance rather than one-off prediction tasks, with a pipeline emphasizing robustness, fairness, and operationalization.

The second item may be used for future continuation of this study – a proposed **Systemic Risk Integration Framework** that translates micro-level predictive outputs into macroprudential indicators (e.g., risk concentration, sectoral heatmaps, and institution-level risk-weighted exposure) for integration into an aggregate vulnerability index. The framework will support early warnings, supervisory dialogues, and targeted mitigation measures during normal periods.

[1] Data is from Kaggle.com: https://www.kaggle.com/datasets/darshandalvi12/non-performing-assets-npa

[2] Systemic Vulnerability Index is a metric that aggregates the different type of risks that threatens the stability of the banking sector.

II. Objectives

The objectives of this study are as follows:

a. Develop and validate an SVM classifier that accurately predicts loan default at the borrower level using standard credit and collateral features, optimized for surveillance.
b. Develop and validate a regularized logistic regression classifier that predicts borrower-level loan default using the same standard credit and collateral features, optimized for surveillance priorities.
c. Compare the results of the two models and recommends which model to deploy.

III. Research Questions

Predictive Performance

How does an SVM-based model perform in predicting borrower-level defaults relative to baseline classifiers (e.g., logistic regression), particularly under class imbalance common in credit datasets?

Moreover, which of the two-machine learning model can effectively predict loan default status by leveraging customer financial and loan-specific features, thereby providing a robust tool for risk assessment in lending decisions?

IV. Significance of the Study

Methodological Significance

Provides a replicable, production-ready approach for borrower-level default prediction using SVM or Logistic Regression, incorporating best practices in preprocessing, class imbalance handling, and model monitoring.

Supervisory Significance

Lays the conceptual groundwork for linking micro-level risk predictions with macroprudential oversight by outlining how borrower-level risk scores could, in future extensions of the study, be aggregated and mapped into systemic indicators. While the empirical demonstration is left for subsequent research, the proposed framework illustrates the potential for integrating such measures into supervisory dashboards and early-warning systems for financial stability monitoring.

Practical Significance

Offers a framework for normal-times surveillance that can be scaled across institutions and loan books, thereby enhancing resilience and reducing the likelihood that localized weaknesses snowball into system-wide crises.

V. Methodology

Utilizing borrower-level data and variables sourced from Kaggle.com, this study aims to construct predictive models for loan default employing machine learning techniques, specifically Support Vector Machines and Logistic Regression. The implementation will be carried out in Python, and the comparative performance of the models will be systematically evaluated. Based on the results, the study will recommend the most effective algorithm to serve as an early warning indicator for loan default risk**.**

The following variables are to be used in the python:

A. Dataset Structure:

Loan_ID, Customer_ID, Loan_Amount, Loan_Type, Credit_Score, Repayment_History, Collateral_Value, Loan_Tenure, Default_Status. Loan_ID and Customer to be dropped in the model.

Please see Annex A for the data sets and Annex B for the Data Dictionary.

Target Variable: Default_Status (binary; mapped to {0,1})
– zero represents non-default and one for default

Feature Groups:

- **Numeric:** Loan_Amount, Credit_Score, Collateral_Value, Loan_Tenure. Repayment_History
- **Categorical:** Loan_Type (e.g., Business, Education, Home, Personal, and Vehicle)

Data Quality Handling:

To maintain the integrity of the dataset, missing values are systematically treated based on their type—numeric or categorical—while issues of class imbalance are appropriately managed to ensure reliable model performance.

- Missing numeric values imputed with median.
- Missing categorical values imputed with the most frequent category. This step in the data preprocessing process involves automatically imputing

missing values in categorical variables with the most frequently occurring category.

- Class imbalance addressed via class_weight="balanced" and threshold calibration.

Preprocessing Pipeline

As part of the data preprocessing stage, the columns *Loan_ID* and *Customer_ID* are excluded from the model, since their presence does not contribute to the prediction of loan defaults.

In essence, preprocess is setting up a comprehensive, automated way to clean and prepare all relevant features for machine learning model, ensuring that each type of data gets the right treatment.

- **Cleaning:** Normalize column names; drop IDs (Loan_ID, Customer_ID).
- **Encoding:** One-hot encode categorical variables with handle_unknown="ignore". This is to turn the categorical variable to numerical columns. The “ignore” for unknown variable will enable the system to ignore should a new category emerge and treat it as example, zero, instead of the system crashing due to undefined item. This makes the code more robust and less likely to break when new, unexpected data comes in.
- **Scaling:** Standardize numeric features (StandardScaler) to ensure SVM stability.

B. Train/Test Split with Stratification

- Stratified split is used with training vs test. The size used in both models (SVM and Logistic Regression) is 70/30 to preserve default prevalence. The data is split into two main parts:

  1. **Training Data: X_train, y_train**. The larger portion (70%) of the dataset that the machine learning model will learn from. It will look for patterns, relationships, and rules within this data.
  2. **Testing Data: X_test, y_test.** This is the smaller (30%), separate portion of the dataset that the model has never seen before. After the model is trained, dataset “test” is to be used to see how well it performs on new, unseen examples. This gives a more realistic idea of the real-world performance.

- For reproducibility, the random_state=42 is used to get the same results every time the code is executed.

- To ensure that the proportion of '0's and '1's in y_train is roughly the same as the proportion of '0's and '1's in y_test, "stratification" is used e.g. stratify=y since the target variable (y) has an imbalanced distribution (e.g., many more '0's than '1's in Default_Status). This ensures both the training and testing sets are representative of the overall data distribution of the target variable.

VI. Modeling: Support Vector Machine (SVM)

In this stage, a Support Vector Machine (SVM) classifier with an RBF kernel was integrated into a preprocessing pipeline to ensure consistent handling of input features. The model was configured with balanced class weights to address class imbalance and probability estimation enabled for downstream risk scoring. Hyperparameter optimization was conducted using a reduced parameter grid — specifically tuning the regularization parameter C {1,2} and the kernel coefficient gamma {{scale},0.1} — to achieve computational efficiency while maintaining robust performance.

Model selection was performed through stratified 3-fold cross-validation, ensuring representative sampling across borrower default classes. The primary evaluation metric was ROC-AUC, chosen for its ability to capture ranking performance in imbalanced datasets, while secondary metrics such as F1-score, Recall, Precision, and PR-AUC provided complementary insights into classification trade-offs.

- **Classifier:** SVC (probability=True, class_weight="balanced", kernel in {linear, rbf}).
- **Hyperparameters:**
  - SVM_C: tune C ∈ {1, 2}, gamma ∈ {"scale", 0.1}
- **Validation:** Stratified K-fold (k=3) with grid search; primary metric **ROC-AUC**; secondary metrics **F1**, **Recall**, **Precision**, and **PR-AUC** (optional).

VII. Evaluation & Diagnostics

The final model was evaluated on the held-out test set to assess its generalization performance. Predictions and probability estimates were generated, and key metrics including Accuracy, Precision, Recall, F1-score, and ROC-AUC were computed. In addition, a detailed classification report was produced to provide class-wise performance insights, ensuring a comprehensive evaluation of the model's effectiveness in predicting loan defaults.

A confusion matrix was plotted to visually assess the classification performance of the SVM on the test set.

- **Metrics:** Accuracy, Precision, Recall, F1, ROC-AUC; and Confusion Matrix.

VIII. Modeling: Logistic Regression

A logistic regression model was trained using the processed dataset with the *lbfgs* solver and a maximum of 1000 iterations to ensure convergence. The model's performance was evaluated on the test set using accuracy and a detailed classification report. These results are presented alongside those of the SVM classifier to enable a direct comparison of predictive effectiveness and to determine which algorithm is more suitable as an early warning indicator for loan default.

- **Max_iter = 1000** - Logistic regression is solved using iterative optimization. This parameter sets the maximum number of iterations the solver can run before stopping.
  - Default is usually 100.
  - Setting it to 1000 ensures the algorithm has enough iterations to converge, especially with larger or complex datasets.
- **Solver = "ibfgs"** This specifies the optimization algorithm used to fit the model.
  - LBFGS stands for Limited-memory Broyden–Fletcher–Goldfarb–Shanno) is a quasi-Newton method.
  - It's efficient for medium to large datasets and supports multinomial logistic regression.
  - It's generally preferred because it's stable and handles regularization well.

IX. Comparative Results of the Two Models

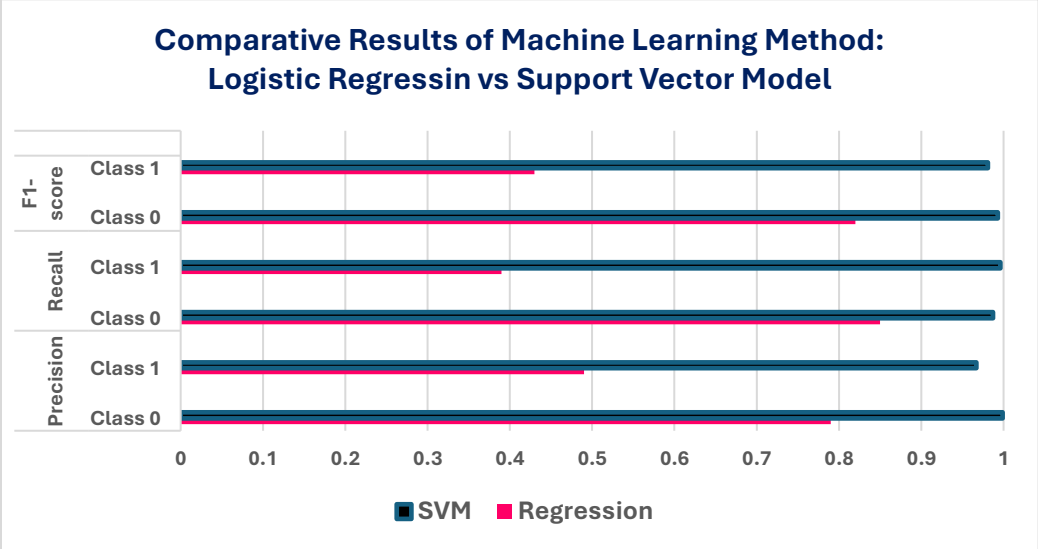


Table 1. Comparing Results of Logistic Regression and Support Vector Model

Metric	Model	Class 0 (Majority)	Class 1 (Minority)	Interpretation
Precision	Regression	0.790	0.490	Class 0: 79% of predicted class 0 instances were correct; Class 1: Less than half of predicted class 1 instances were correct.
	SVM	0.999	0.967	Class 0: Nearly perfect precision; almost all predicted class 0 instances were correct; Class 1: Very high precision; most predicted class 1 instances were correct.
Recall	Regression	0.850	0.390	Class 0: 85% of actual class 0 instances were correctly identified; Class 1: Only 39% of actual class 1 instances were identified
	SVM	0.987	0.996	Class 0: Almost all actual class 0 instances were captured; Class 1: Nearly all actual class 1 instances were correctly predicted
F1-score	Regression	0.820	0.430	Class 0: Good balance between precision and recall; Class 1: Weak performance.
	SVM	0.993	0.981	Class 0: Excellent performance; Class 1: Very strong performance
Support	Regression	65,665	24,402	Class 0: 65,665 samples; Class 1: 24,402 samples
	SVM	65,665	24,402	

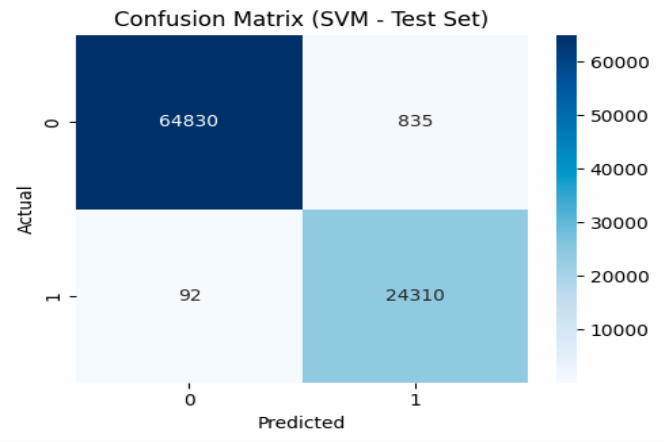
The comparative evaluation highlights clear differences between Logistic Regression and SVM. Logistic Regression demonstrates reasonable performance for the majority class (precision 0.79, recall 0.85, F1-score 0.82) but struggles with the minority class (precision 0.49, recall 0.39, F1-score 0.43). In contrast, the SVM achieves near-perfect results across both classes, with precision, recall, and F1-scores all exceeding 0.96. This indicates that while Logistic Regression provides a baseline model, the SVM offers substantially stronger predictive capability and balanced performance, making it more effective for loan default detection in both majority and minority borrower groups.

- **Logistic Regression**
    - Performs reasonably well for the majority class but struggles with the minority (defaults).

- Precision and recall for default borrowers are relatively low (≈0.49 and 0.39), meaning many defaults are missed.
- Advantage: simplicity, interpretability, and ease of deployment at scale.

- **Support Vector Machine (SVM)**
  - Achieves near-perfect precision and recall across both majority and minority classes (>0.96).
  - Strong F1-scores indicate balanced performance.
  - Advantage: highly accurate, robust to imbalance, and reliable in capturing rare default events.

X. Resulting Confusion Matrix Using SVM only



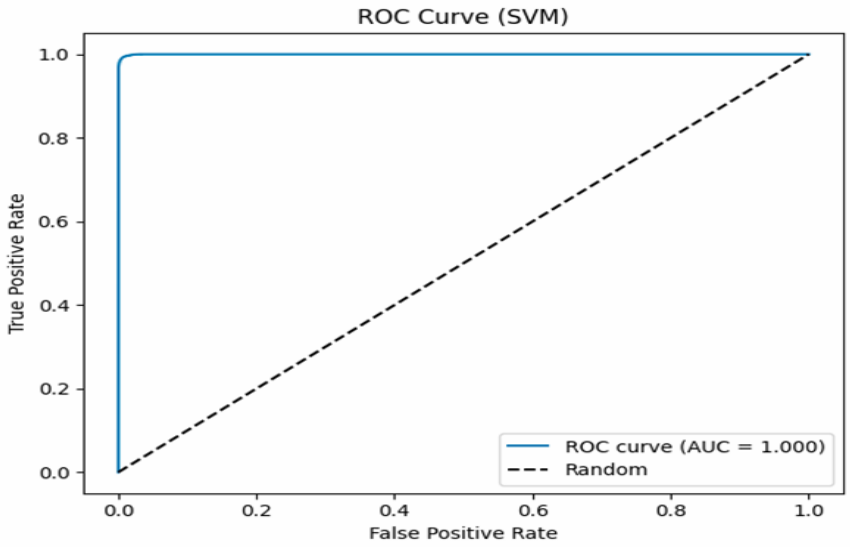
The confusion matrix summarizes the out-of-sample classification performance of the SVM model on the test set by comparing predicted outcomes with actual borrower default status.

- **True Negatives (64,830):** The model correctly identifies a large majority of non-defaulting borrowers. This indicates strong performance in recognizing low-risk borrowers.
- **False Positives (835):** A relatively small number of non-defaulting borrowers are incorrectly classified as defaulters. While this represents some overestimation of risk, the magnitude is limited compared to the total number of non-default cases.
- **False Negatives (92):** Only a very small number of actual defaulters are misclassified as non-defaulters. This is particularly important from a financial stability and risk management perspective, as missed defaults are typically more costly.

- **True Positives (24,310):** The model successfully identifies most actual defaulters, demonstrating strong discriminatory power for high-risk borrowers.

Overall, the results suggest that the SVM model exhibits **high classification accuracy and strong default-detection capability**, with a particularly low rate of false negatives. This implies that the model is effective as an early-warning tool, prioritizing the identification of risky borrowers while maintaining a low level of misclassification among safe borrowers—an outcome that is well aligned with supervisory and macroprudential risk-monitoring objectives.

XI. ROC Curve Plot of SVM only



The ROC curve illustrates the SVM model's ability to discriminate between defaulting and non-defaulting borrowers across all possible classification thresholds.

The curve lies very close to the upper-left corner of the plot, indicating that the model achieves a high true positive rate while maintaining an extremely low false positive rate over a wide range of thresholds. This reflects strong separability between the two classes.

The **Area Under the Curve (AUC) of 1.000** signifies *near-perfect discriminatory power* on the test set. In practical terms, this means that the model almost always assigns a higher predicted risk score to a defaulting borrower than to a non-defaulting one.

From a credit risk and supervisory perspective, this result suggests that the SVM model is highly effective as an early-warning tool, capable of ranking borrowers by

default risk with exceptional accuracy. However, such a near-perfect AUC also warrants caution: it may reflect favorable data characteristics, strong feature informativeness, or potential issues such as data leakage or overfitting. As such, robustness checks—such as cross-validation, out-of-time testing, or stress scenarios—would be important in subsequent analysis to confirm the model's stability and real-world applicability.

XII. Recommendation for Early Warning Indicator

Given the large dataset size (tens of thousands of samples per class) and the critical need to detect minority events (defaults), the **SVM model is highly recommended for deployment** as the early warning indicator.

- **Why SVM?**
    - Superior ability to capture minority class defaults without sacrificing majority class accuracy.
    - High recall ensures few defaults are missed — essential for surveillance and systemic risk prevention.
    - High precision reduces false alarms, making monitoring more efficient.

XIII. Role in Surveillance Tools

- During **normal times**, the SVM can serve as a **continuous monitoring tool**, flagging borrowers with elevated default risk.
- Its calibrated probability outputs can be integrated into dashboards for **risk surveillance**, allowing regulators or institutions to track shifts in borrower risk profiles.
- By detecting early signals of rising default probabilities, the model supports **systemic risk prevention** by enabling proactive interventions before defaults accumulate into broader financial instability
- Deploy the SVM classifier as the **primary early warning indicator**, supported by Logistic Regression as a benchmark or secondary model for interpretability. This dual setup balances accuracy (SVM) with transparency (Logistic Regression), ensuring robust surveillance during normal times and enhancing systemic risk prevention strategies.