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

By: Jessica P. Hutalla

In Completion of the Asian Institute of Management and Emeritus Post Graduate Diploma on Artificial Intelligence and Machine Learning

Objective

Modeling Objective:

- Develop and validate a Support Vector Model classifier that accurately predicts loan default at the borrower level using standard credit and collateral features, optimized for surveillance.
- 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.
- Compare the results of the two models and recommends which model to deploy

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?

Significance of the Study

Methodological Significance

- Provides a replicable, production-ready approach for borrower-level default prediction using SVM and 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.

Methodology

Utilizing borrower-level data and variables sourced from Kaggle.com, the implementation will be carried out in Python, and the comparative performance of the models will be systematically evaluated. Stratified split will be used with training vs test of 70/30 to preserve default prevalence. Based on the results, the study will recommend the most effective algorithm to serve as an early warning indicator for loan default risk.

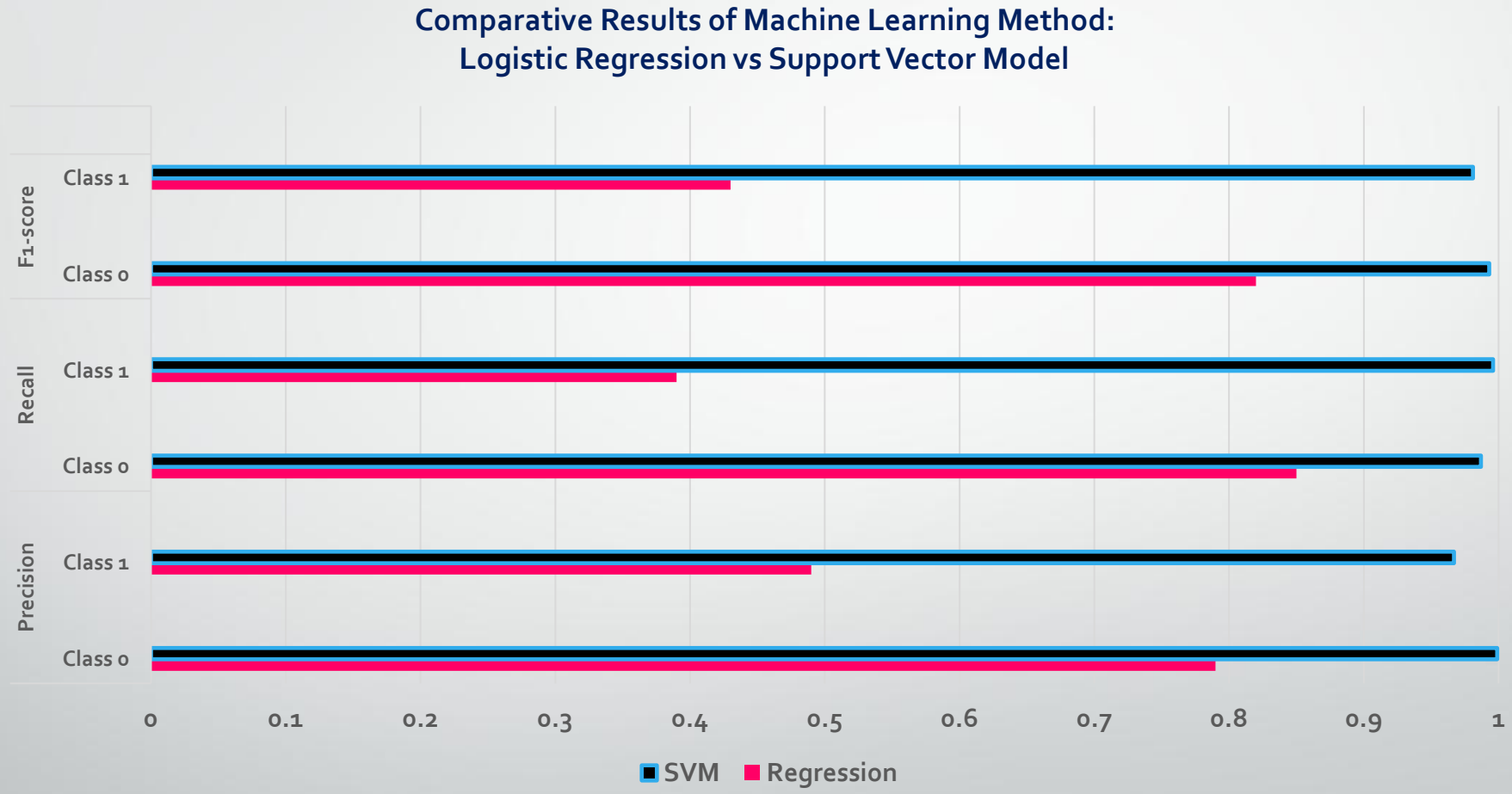
Modeling: Support Vector Machine (SVM)

A Support Vector Machine (SVM) classifier with an RBF kernel will be integrated into a preprocessing pipeline to ensure consistent handling of input features.

Modeling: Logistic Regression

A logistic regression model will be 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.

Results



Result Discussion

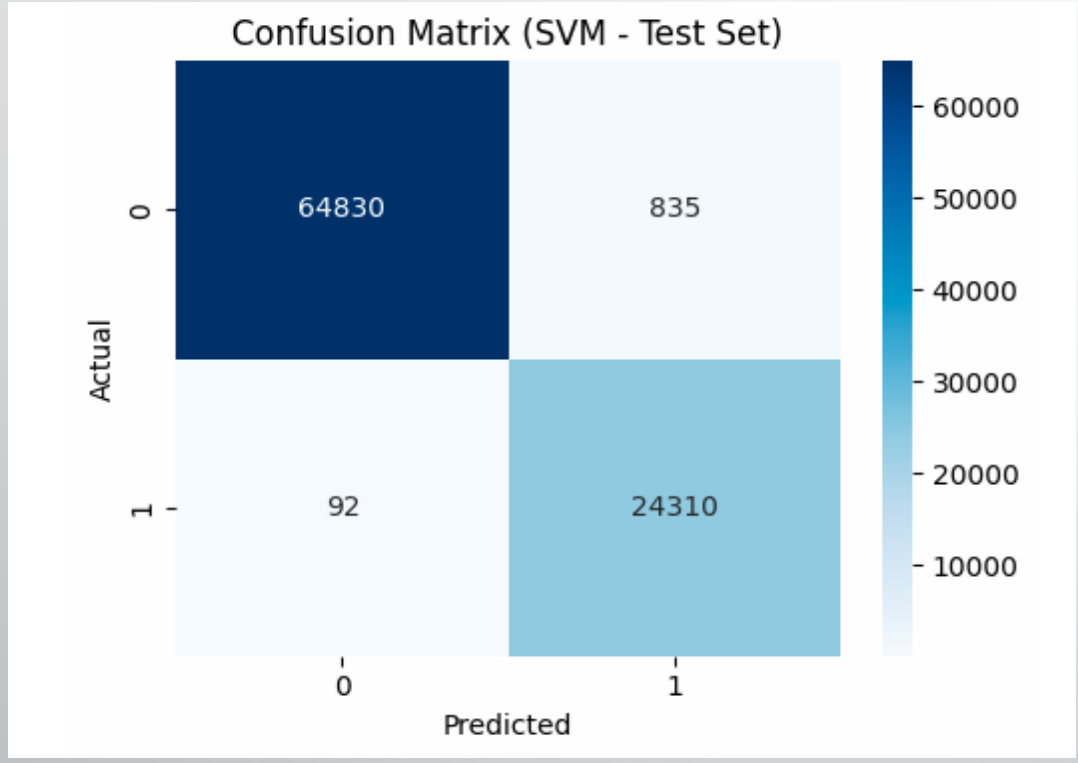
Logistic Regression

- 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.

SVM Resulting Confusion Matrix



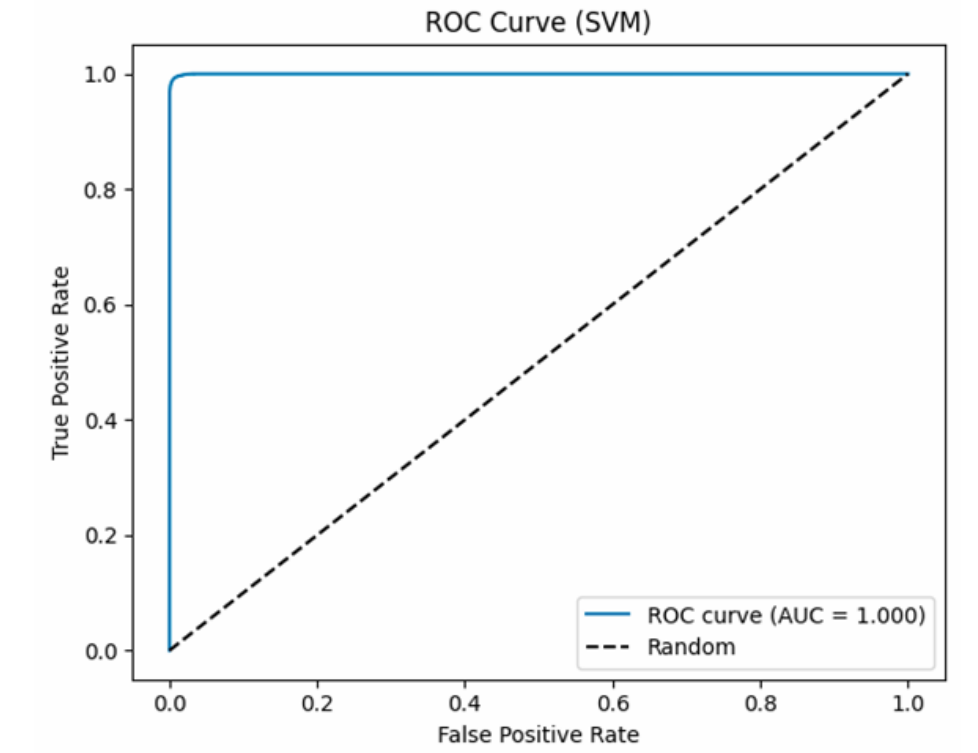
True Negatives (64,830): The model correctly identifies a large majority of non-defaulting borrowers. This indicates strong performance in recognizing low-risk borrowers.

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.

SVM ROC-AUC Curve



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.

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 for systemic risk.

- 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.

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.

End of the Report