ESILV — École Supérieure d’Ingénieurs Léonard de Vinci

Course: Machine Learning

Predicting Systemic Risk in Rural Financial Institutions

Guillaume Xu
Hugo Vignet

Professor: Marius Ortega

December 10, 2025

Contents

1 Introduction and Business Case

The stability of financial systems is a cornerstone of economic health, particularly in rural sectors where institutions operate under unique constraints. Unlike their urban counterparts, rural financial institutions face a high degree of economic uncertainty driven by specific local factors such as agricultural performance, susceptibility to natural disasters, and fluctuating liquidity constraints.

A failure in one such institution can trigger contagion effects, destabilizing the broader rural economy. Consequently, the ability to accurately assess the Systemic Risk Level of these institutions is critical for maintaining financial stability.

This project aims to develop a predictive machine learning framework capable of classifying rural financial institutions into three distinct risk categories: Low, Medium, and High. By leveraging a combination of internal financial metrics and external macroeconomic indicators, the objective is to deploy an automated early warning system.

This initiative is intrinsically linked to the field of Actuarial Science, specifically within the domains of Quantitative Risk Management and Solvency Analysis. Actuaries are tasked with evaluating the probability of future adverse events and financial insolvency. By utilizing historical data to model future risk states, this project applies core actuarial principles to support regulatory supervision and capital requirement assessments, akin to the pillars of the Solvency II framework.

2 Data Description and Exploratory Analysis

The analysis is based on the ”Rural Financial Systemic Risk Dataset”, a synthetic representation of the financial health and operating environment of rural banks. The dataset comprises 10,293 observations and 21 variables, providing a comprehensive snapshot of the sector. The feature set includes critical financial indicators such as the Capital Asset Pricing Model, Equities, Loan Amounts, and Liquidity Ratio, alongside macroeconomic variables like Agricultural Performance, Inflation Rate, and Natural Disasters.

The target variable, Systemic_Risk_Level, categorizes institutions into Low (0), Medium (1), or High (2) risk. The initial exploratory analysis revealed that the data was clean, with no missing values or duplicates. However, a significant challenge emerged in the distribution of the target variable. The classes were imbalanced: while the Medium and High-risk categories contained approximately 4,117 observations each, the Low-Risk category was underrepresented with only about 2,000 observations. This imbalance poses a risk of bias in standard classification algorithms, which may favor majority classes.

Furthermore, correlation analysis highlighted strong relationships between specific risk factors, such as Risk_Factor_Equities, Capital Ratio, Liquidity Ratio and the target variable.

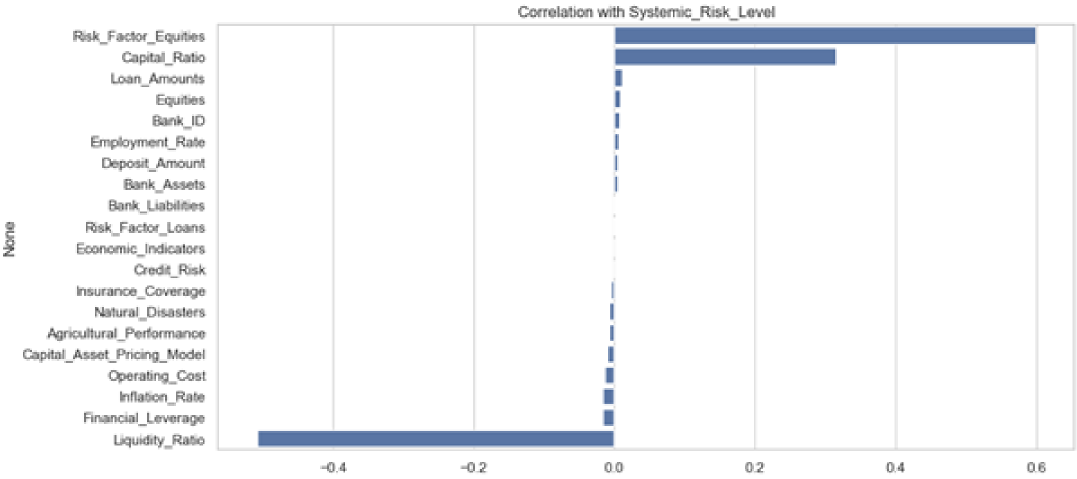


Figure 1: Correlation with Systemic Risk Level

Boxplot visualizations confirmed that features like the Liquidity Ratio and Capital Ratio exhibit clear distributional distinctions across risk levels, suggesting high predictive potential.

The boxplots indicate that the dataset does not provide meaningful statistical separation across the three risk classes: Low, Medium, and High. Although some features exhibit logical directional trends, all variables show substantial overlap in their interquartile ranges, preventing the formation of clear decision boundaries. The strongest overlap is observed between the Medium and High risk groups, whose distributions are nearly indistinguishable for many variables. This leads models to frequently confuse these two classes, a pattern later confirmed in the confusion matrices.

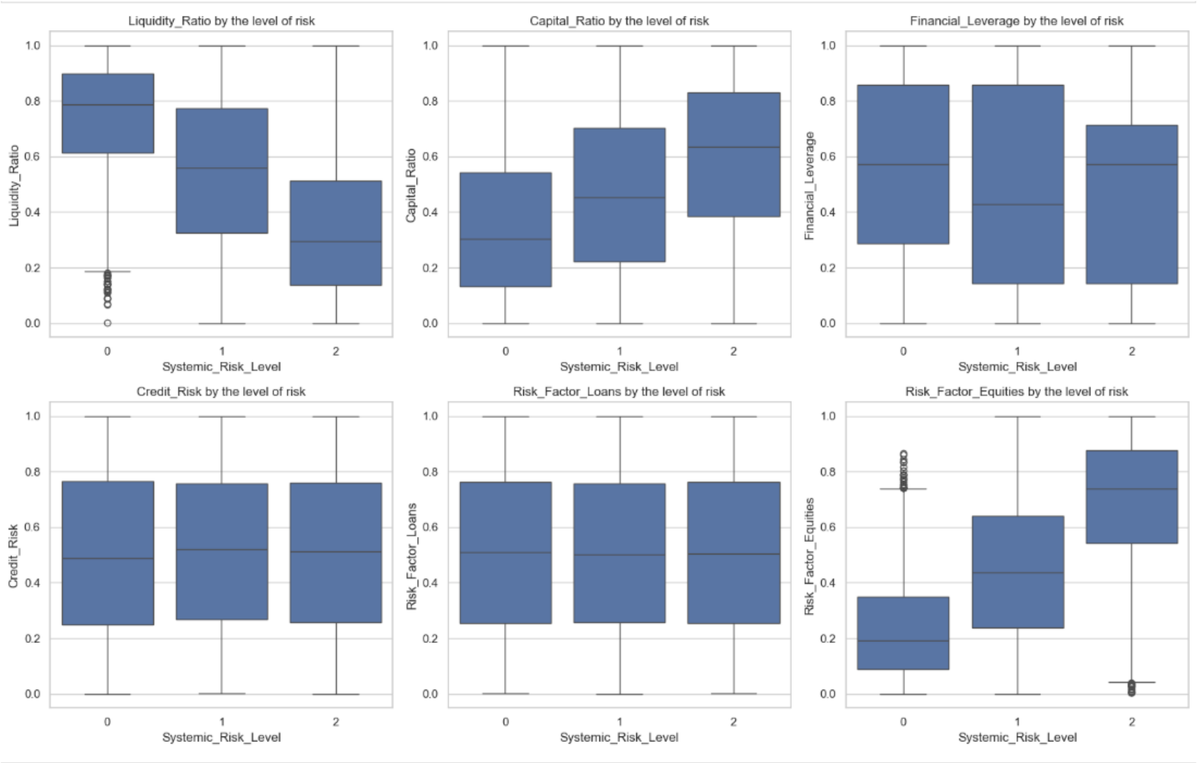


Figure 2: Boxplots of Key Features by Risk Level (3 Classes)

Several features, including Credit Risk and Loans, offer minimal discriminatory value and contribute noise rather than useful signal. This structural lack of separability directly explains the recurring recall problems in the multiclass models, particularly the frequent misclassification of Medium-risk institutions as Low-risk. Because the dataset does not naturally form distinct clusters, even advanced algorithms such as Random Forest are unable to reliably differentiate among the three risk levels. So, the boxplot analysis shows that the dataset does not support an effective multiclass classification framework.

This validated the decision to transition to a binary classification setup, enabling the model to focus on a single, more meaningful decision boundary: Risk versus No Risk.

3 Methodology and Modeling Strategy

3.1 Multiclass Approach

The initial problem was formulated as a multiclass supervised classification task. Two baseline models were then developed to establish foundational performance benchmarks:

- **Multinomial Logistic Regression:** This linear model achieved moderate accuracy but was limited by the presence of non-linear relationships within the dataset.

- **Random Forest Classifier:** As a non-linear ensemble method, the Random Forest model captured more complex patterns and demonstrated stronger predictive performance compared to the linear baseline.

Logistic Regression accuracy was approximately 83 percent, while the Random Forest model achieved roughly 92 percent accuracy and a ROC-AUC of 0.97. Although these metrics suggest high performance, a deeper evaluation revealed insufficient recall, particularly for Medium and High risk institutions. This issue is especially consequential in systemic risk assessment, where failing to correctly classify a high-risk institution can have significant implications. The primary causes of poor recall were attributed to insufficient class separability and imbalanced class distributions. These structural limitations indicated that the multiclass framework could not consistently support reliable risk-based decision-making.

3.2 Why 2 Classes?

The multiclass modeling phase offered valuable preliminary insights but ultimately revealed fundamental limitations that prevented the accurate and reliable prediction of systemic risk categories. Although initial models such as Multinomial Logistic Regression and Random Forest achieved relatively high accuracy scores, their recall performance remained insufficient, particularly for distinguishing Medium and High risk institutions.

This weakness was directly tied to the statistical structure of the dataset: significant overlap among key features, especially between the Medium and High risk classes, created blurred decision boundaries that persisted across all modeling approaches. A deeper examination of the dataset highlighted a second, equally important motivation for abandoning the multiclass framework. The class distribution was inherently imbalanced, with the Low-risk category representing the minority class and the Medium- and High-risk categories collectively forming a dominant majority. This imbalance limited the model's ability to learn stable and generalizable patterns for the smaller class, resulting in high variance and reduced reliability.

When the Medium and High categories were merged into a single Risk class, the distribution became considerably more balanced and statistically well-structured. This restructuring offered several methodological advantages:

1. The larger and more cohesive Risk class allowed the model to benefit from the Law of Large Numbers. With a higher volume of observations supporting a single class, the learned patterns became more stable, reducing variance and improving

the model's capacity to generalize.

2. In high-dimensional feature spaces, class boundaries become more meaningful when the underlying distributions are sufficiently dense. The binary configuration strengthened the estimation of these boundaries, making them clearer and more tractable than in the multiclass scenario, where overlapping distributions obscured separability.

In contrast, the multiclass configuration suffered from limited inter-class separation and inadequate representation of the Low-risk group, leading to unstable decision surfaces and inconsistent predictive outputs. By consolidating the Medium and High categories, the modeling task shifted from trying to resolve subtle and statistically unsupported distinctions to focusing on a single, operationally relevant question: identifying whether an institution poses systemic risk or not.

Overall, the combination of structural data characteristics, performance limitations, and theoretical considerations supported the transition to a binary classification framework. This approach enhances statistical power, reduces variance, improves recall, and ultimately provides a more reliable basis for systemic risk detection and supervisory decision-making.

3.3 Exploratory Analysis of Two Classes

The boxplot analysis shows that the dataset does not provide meaningful statistical separation among the three original risk classes: Low, Medium, and High. Most financial variables display substantial overlap in their interquartile ranges, with the strongest overlap occurring between the Medium- and High-risk groups. This lack of separability prevents models from learning stable decision boundaries and leads to frequent misclassification, especially the tendency to classify Medium-risk institutions as Low-risk.

When the Medium- and High-risk categories are merged into a single Risk class, a clearer pattern emerges. The resulting distribution becomes denser and more homogeneous, improving estimator stability through the Law of Large Numbers.

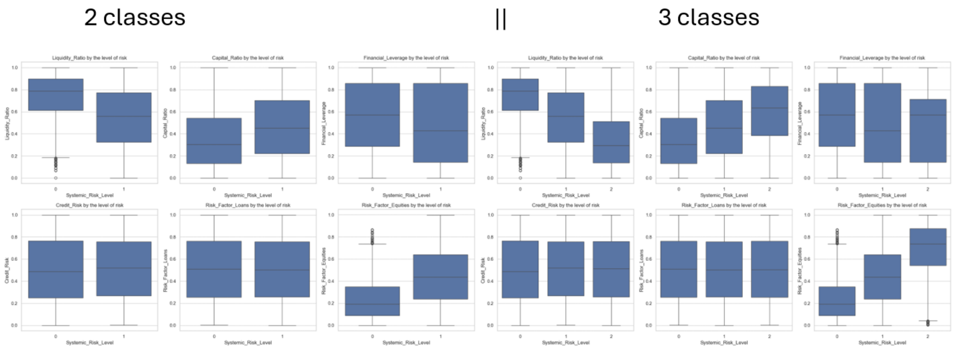


Figure 3: Comparison: 3 Classes (Multiclass) vs. 2 Classes (Binary)

The binary boxplots show a visibly sharper separation between the No Risk and Risk groups, allowing models to form more robust and interpretable decision boundaries. This binary structure also reduces the likelihood of predicting No Risk for institutions that already exhibit early signs of vulnerability.

Overall, the structural characteristics of the dataset demonstrate that a multiclass framework cannot provide reliable systemic risk classification. Transitioning to a binary model enhances statistical power, improves recall, stabilizes estimators, and aligns the analysis with the practical goal of determining whether an institution is risky or not.

3.4 Binary Modeling Strategy

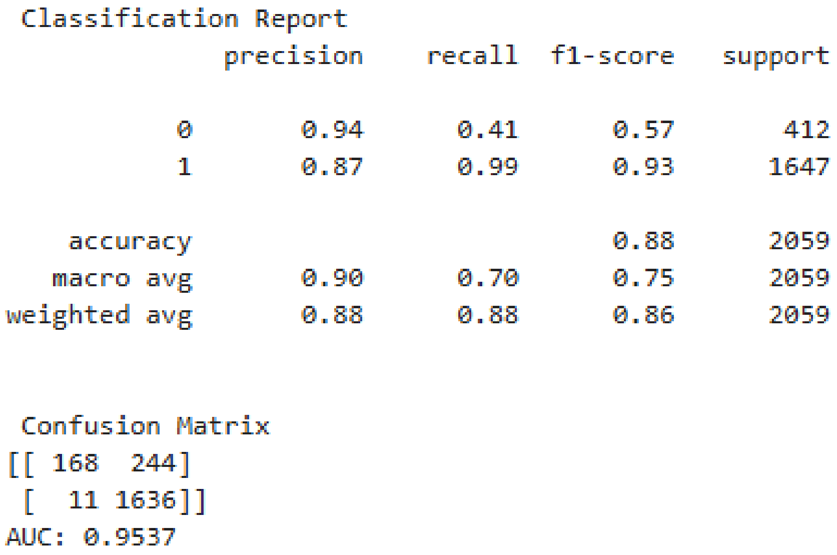
Three modeling approaches were explored to address the systemic risk prediction problem. The Multinomial Logistic Regression model served as a linear baseline, while the Random Forest Classifier was introduced to capture nonlinear relationships and improve overall robustness. However, persistent class overlap and limited separability in the multiclass framework ultimately motivated a shift toward a Binary Classification formulation distinguishing Risk from No Risk.

In the binary setting, class imbalance remained a central challenge, with the No Risk class significantly underrepresented. To correct this imbalance and stabilize decision boundaries, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE increased the representation of safe institutions, improved class symmetry, and enhanced the model's ability to detect subtle differences between the two categories without compromising overall performance.

Following SMOTE oversampling and feature standardization, several algorithms were evaluated, and XGBoost emerged as the most effective model. Its ability to capture nonlinear interactions, mitigate the effects of noisy features, and adapt to imbalanced data made it particularly well suited to the risk detection task. Hyperparameter tuning using GridSearchCV was conducted with Recall chosen as the optimization metric, reflecting the supervisory priority of minimizing false negatives, the most dangerous errors in systemic risk monitoring.

4 Results and Discussion

The optimized XGBoost model demonstrated exceptional sensitivity, achieving a Recall of 0.9933 and missing only 11 high-risk institutions out of 1,647.



```
 Classification Report
              precision    recall  f1-score   support

           0       0.94      0.41      0.57       412
           1       0.87      0.99      0.93      1647

    accuracy                           0.88      2059
   macro avg       0.90      0.70      0.75      2059
weighted avg       0.88      0.88      0.86      2059


 Confusion Matrix
[[ 168  244]
 [  11 1636]]
AUC: 0.9537
```

Figure 4: Classification Report and Confusion Matrix (XGBoost Binary)

Although this performance resulted in 244 false positives, such outcomes are acceptable in a regulatory context, where precautionary investigation is preferable to overlooking emerging vulnerabilities.

Feature importance analysis further revealed that Risk_Factor_Equities, Liquidity_Ratio, and Capital_Ratio were the most influential drivers of systemic risk.

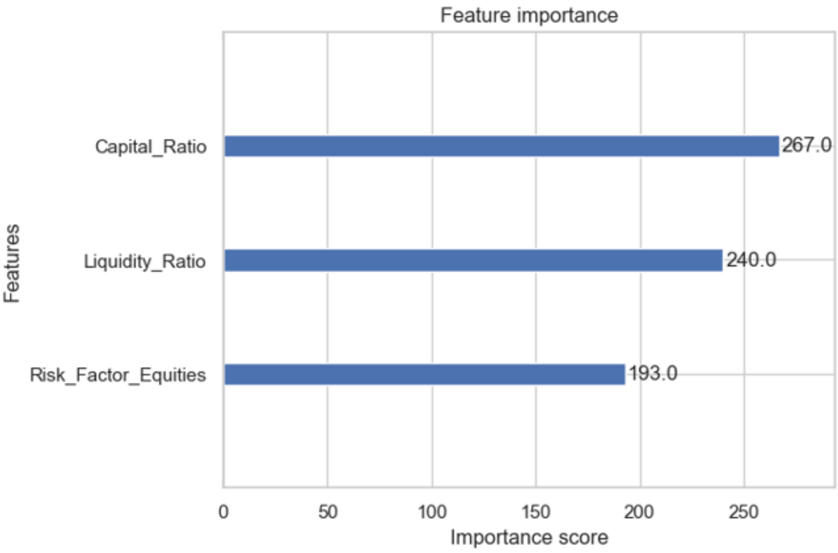


Figure 5: Feature Importance Analysis

Overall, the modeling pipeline successfully transitioned from baseline multiclass approaches to a high-recall binary framework capable of functioning as a reliable early warning system for supervisory decision-making.

5 Conclusion

The final binary modeling pipeline, integrating SMOTE oversampling, standardized feature scaling, recall-based optimization, and a tuned XGBoost classifier, produced a robust and supervision-oriented predictive framework.

By increasing the density and representativeness of the minority class, SMOTE contributed directly to improving estimator stability and reducing bias, enabling the model to form clearer and more reliable decision boundaries. The high Recall performance of the optimized XGBoost model confirms its effectiveness as an early warning tool for systemic risk detection. Its capacity to identify institutions at elevated risk, even at the cost of additional false positives, aligns closely with actuarial and regulatory priorities, where precautionary alerts are preferable to undetected vulnerabilities.

The model's feature importance results further reinforce its interpretability, highlighting clear and economically meaningful indicators of systemic fragility. Overall, the resulting methodology delivers a strong, high-recall predictive system capable of supporting proactive supervisory decision-making and providing a dependable analytic foundation for systemic risk monitoring.