WeTrust Credit Scoring Model

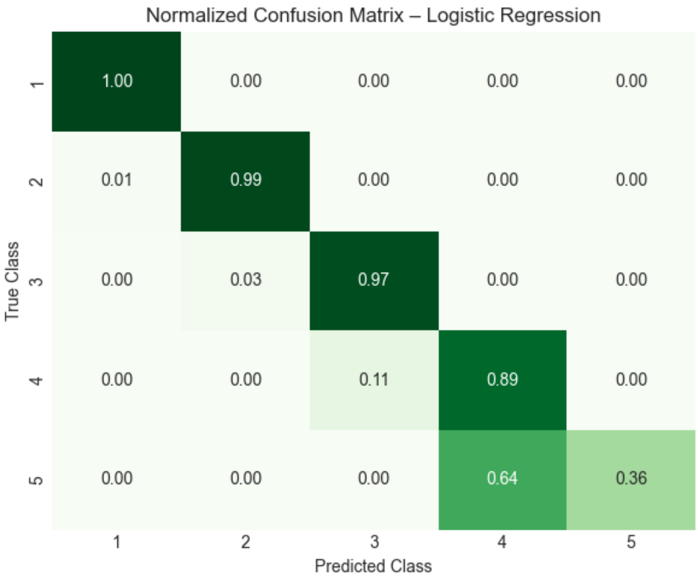
Model Evaluation and Interpretation

1. Overview

This report presents the main visual diagnostics used to evaluate and interpret the performance of the **WeTrust Credit Scoring Model**. The objective of the model is to predict the *creditworthiness class* of an individual, assigning each person to a discrete merit category ranging from 1 to 5. Class 1 corresponds to a user with low credit reliability (high financial risk), whereas class 5 identifies a highly trustworthy profile, eligible for microcredit loan approval.

The following figures provide complementary insights into model behavior: the confusion matrix illustrates classification accuracy across these merit classes, the feature importance analysis identifies the key predictors driving the credit score, and the ROC–PR curves assess the model's probabilistic discrimination capability.

1.1. Normalized Confusion Matrix



	1	2	3	4	5
1	1.00	0.00	0.00	0.00	0.00
2	0.01	0.99	0.00	0.00	0.00
3	0.00	0.03	0.97	0.00	0.00
4	0.00	0.00	0.11	0.89	0.00
5	0.00	0.00	0.00	0.64	0.36

Figure 1: Normalized Confusion Matrix for the Logistic Regression model.

The normalized confusion matrix summarizes the relationship between *true classes* (on the Y-axis) and *predicted classes* (on the X-axis). Each cell (i, j) represents the proportion of instances of class i predicted as class j. The matrix above is row-normalized so that each row sums to 1.00. Diagonal elements correspond to correct classifications, while off-diagonal elements quantify misclassifications.

Formally:

$$M_{i,j} = \frac{N_{i,j}}{\sum_k N_{i,k}}$$

where $N_{i,j}$ is the number of samples of true class i predicted as class j. High diagonal values (close to 1) indicate strong predictive reliability across merit classes.

It is important to note that the model has been deliberately parameterized to adopt a **conservative decision policy** regarding microcredit eligibility. In practical terms, when uncertain, the model is designed to favor assigning an individual to a *lower* merit class rather than a higher one. This bias minimizes the risk of granting credit to potentially unreliable applicants, at the cost of occasionally underestimating the creditworthiness of some users.

1.2. Feature Importance Analysis

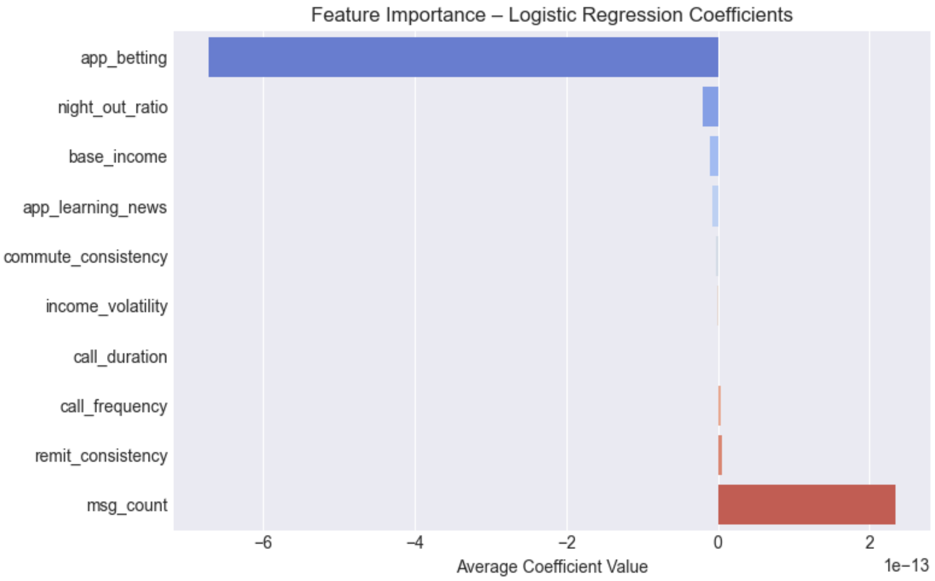


Figure 2: Feature Importance — Logistic Regression Coefficients.

Feature importance quantifies how much each input variable contributes to the model's predictions. In a Logistic Regression classifier, each feature x_j is associated with a coefficient β_j in the model:

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$$

where P is the predicted probability of belonging to a higher merit class.

A positive coefficient ($\beta_j > 0$) increases the log-odds of being in a higher class, while a negative coefficient reduces it. The plot ranks features by the magnitude and sign of their average coefficients: blue bars represent negative effects (features that decrease predicted credit reliability), and red bars represent positive effects. For example, variables such as *app_betting* show a strong negative coefficient, indicating a lower creditworthiness, while the characteristics of regularity behavioral (*msg_count*, *remit_consistency*) have a positive impact.

1.3. ROC and Precision–Recall Curves

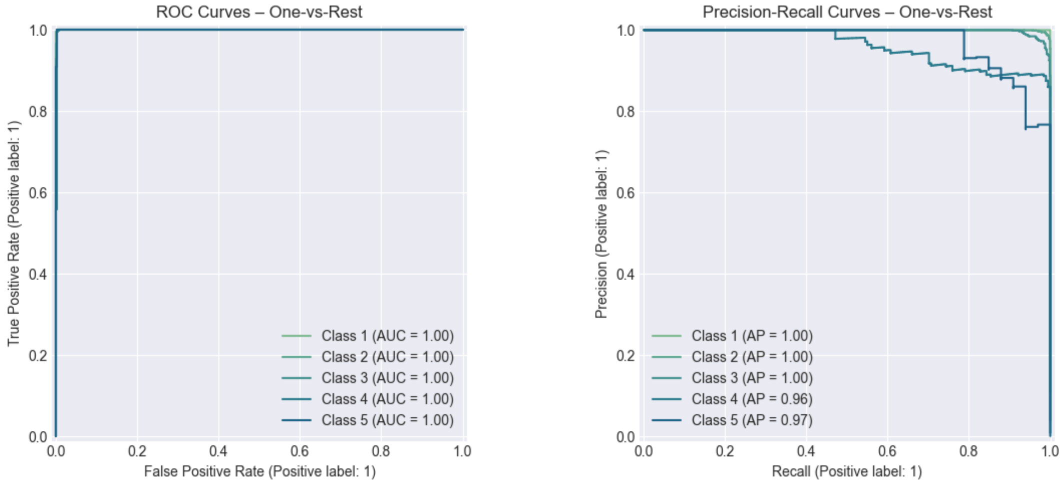


Figure 3: ROC and Precision–Recall curves for each merit class (One-vs-Rest evaluation).

The left panel displays the **Receiver Operating Characteristic (ROC)** curves, illustrating the trade-off between *True Positive Rate* (TPR) and *False Positive Rate* (FPR) across probability thresholds. Each curve corresponds to one merit class in a one-vs-rest setting. The *Area Under the Curve (AUC)* quantifies overall separability: AUC = 1.0 indicates perfect discrimination, whereas 0.5 corresponds to random guessing. Values near 1.0 confirm excellent classification ability across all classes.

The right panel presents the **Precision–Recall (PR)** curves, which show how *precision* (the proportion of correctly predicted positives) varies with *recall* (the proportion of actual positives correctly identified). The *Average Precision (AP)*—the area under each PR curve—is especially relevant for imbalanced datasets. AP values above 0.8 indicate high precision even at high recall levels. Overall, the combination of high AUC and AP scores confirms that the model not only classifies accurately but also produces well-calibrated probabilities, an essential property for credit scoring applications.

Disclaimer on Model Performance

The performance metrics and diagnostic plots presented in this report should be interpreted in light of the dataset size used during model training and evaluation. As with any statistical or machine learning model, predictive accuracy and stability are strongly dependent on the amount and diversity of available data. To assess this sensitivity, simulations were conducted with datasets of increasing size — approximately 1,000, 10,000, and 100,000 samples.

Results indicate that larger datasets lead to more stable coefficient estimates, smoother probabilistic calibration, and improved generalization. Conversely, with limited data the model may exhibit higher variance and reduced reliability in class boundaries. Therefore, the reported figures should be considered indicative of expected behavior under the tested data conditions, rather than absolute performance guarantees.