

Report

- **Logistic Regression Model**

Model performance was evaluated on the validation set. The LR model achieved an accuracy of **93.9%** and an AUC of **0.82**. Although the accuracy is high due to the strong class imbalance, the AUC is a more reliable metric because it measures the model's ability to distinguish between default and non-default customers. The LR model identified non-default customers well but captured only a moderate share of true defaults.

Positive coefficients indicate higher default risk, and the strongest positive predictors include: **uti_card**; **non_mtg_acc_past_due_12_months_num**; **avg_card_debt**; **inq_12_month_num**; **mortgages_past_due_6_months_num**. These variables reflect high credit utilization, recent delinquencies, and increased credit-seeking behavior—all consistent with classical credit risk theory.

The strongest negative coefficients (lower default risk) include: **tot_credit_debt**; **credit_age**, **card_age**, and **credit_good_age** (longer credit history); **rep_income** (higher income); **ind_acc_XYZ** (existing customer of the institution).

Overall, logistic regression provides interpretability and clear directional relationships, making it a good baseline model. However, because LR assumes linear decision boundaries, it is limited in capturing nonlinear interactions typical in credit default behavior. This motivates the use of more flexible machine learning models.

- **Random Forest Model**

For the machine learning approach, I developed a **Random Forest (RF)** model. Random Forests are well-suited for tabular credit-risk data because they capture nonlinear relationships, interactions among predictors, and are robust to noisy features.

I first trained a baseline RF model using default hyperparameters (with `class_weight="balanced"` to mitigate class imbalance). The baseline RF achieved a validation accuracy of **92.2%** and an AUC of **0.86**, already outperforming logistic regression, particularly in its ability to identify defaulting customers. The recall for the default class was **significantly higher** than that of LR, demonstrating RF's superior sensitivity to risky customers.

To further improve performance, I conducted a **lightweight hyperparameter search** over five reasonable configurations, varying the number of trees (`n_estimators`), tree depth (`max_depth`), and minimum leaf size (`min_samples_leaf`). Each candidate model was trained on the training set and evaluated on the validation set using **AUC as the primary metric**, as it is most appropriate for imbalanced credit-risk problems.

The best configuration was identified as: **`n_estimators = 400`, `max_depth = 8`, `min_samples_leaf = 5`, Validation AUC = 0.8593**.

Although the AUC improvement over the baseline RF is modest, this tuned model consistently exhibits strong generalization and competitive performance. The reduced tree depth (8) mitigates overfitting compared with deeper trees, while a leaf size of 5 stabilizes the model without oversimplification.

Given its higher AUC, improved recall for default cases, and ability to model nonlinearities, the **tuned Random Forest** was selected as the preferred machine learning model. For the bonus competition and hidden-test evaluation, this tuned model was **retrained on all available labeled data** (training + validation) to maximize predictive power.

- **Compare LR and RF model & select one for credit**

The logistic regression (LR) model provides clear interpretability and highlights meaningful risk drivers such as utilization, credit age, and delinquency history. However, its predictive power is limited: LR has a lower AUC (~0.82) and identifies fewer true defaults due to its linear structure.

The tuned Random Forest (RF) model performs significantly better, achieving an AUC of **0.86**, higher recall for default cases, and stronger discrimination between risky and non-risky customers. Although less interpretable, RF captures nonlinear patterns and interactions that LR cannot model.

For credit approval, I select the **tuned Random Forest** because it more effectively detects high-risk applicants, which is critical in minimizing losses. The performance improvement outweighs the loss of interpretability.

- **What performance metrics you chose and why**

I use **AUC**, **recall for the default class**, and **accuracy**.

AUC is the primary metric because the dataset is imbalanced, and AUC measures ranking and discrimination rather than raw accuracy.

Recall for default (class 1) is essential because missing a high-risk applicant is far more costly than incorrectly flagging a safe customer.

Accuracy is reported but not emphasized, since it can be misleading in imbalanced datasets.

- **How to make decisions on credit card applications**

The model outputs a predicted probability of default. I would use this probability to define decision rules:

Applicants with **low predicted risk** (e.g., probability < 0.10) are **approved automatically**.

Applicants with **moderate risk** (e.g., 0.10–0.25) undergo **manual review** or require additional documentation.

Applicants with **high predicted risk** (e.g., > 0.25) are **declined**.

These thresholds can be calibrated based on the bank's risk appetite and loss tolerance.

- **Do customers who already have an account with the financial institution receive any favorable treatment in your model?**

Yes, the logistic regression results show that `ind_acc_XYZ` has a **negative coefficient**, meaning existing customers of the institution have a lower predicted default probability, holding other variables constant. Random Forest feature importance also indicates that this variable contributes to risk assessment. Therefore, existing customers appear to receive slightly more favorable treatment because historically they default less.