Credit Card Default Risk: One-Month PD Modelling and Interpretation

1 Problem

Credit card issuers face the problem of assessing the probability that a client will default on their payments in the near future. From a business perspective, the bank needs credible, calibrated probability of default (PD) to quantify expected losses, set credit limits, and decide when to intervene (e.g. collection actions, limit reductions). These decisions depend not only on ranking (discrimination) but also on how probabilities are scaled and on the cost of missing a default versus blocking a good client. For this reason, our project is organized around three pillars: (i) building well-calibrated PD models; (ii) linking them to cost-based global and segment-specific thresholds; (iii) interpreting the behavioral drivers behind high PDs. The response variable for our project is:

$$Y = \texttt{default\_payment\_next\_month} \in \{0, 1\}.$$

2 Data

We use the public UCI dataset "Default of credit card clients in Taiwan" (April–September 2005). The original data contain 30 000 credit card holders and 25 variables, including socio-demographic characteristics (sex, education, marital status, age), credit limit, repayment status in the last six months (`PAY_0`–`PAY_6`), bill statements (`BILL_AMT1`–`BILL_AMT6`) and amounts paid for each of the last six months (`PAY_AMT1`–`PAY_AMT6`), plus the binary target `default_payment_next_month`, with a default prevalence of approximately 22%, so the problem is moderately imbalanced and accuracy at threshold 0.5 is not an informative metric.

Data cleaning proceeds as follows: we read the Excel file robustly without trusting the original header, remove completely empty columns, drop the first two bad rows (empty or header-like), assign the 25 official UCI column names, convert all columns to numeric, remove the technical ID and create an internal row identifier. We recode `EDUCATION` by grouping values {0, 5, 6} into "others" (4), and `MARRIAGE` by mapping 0 to "others" (3), yielding coherent category sets aligned with the documentation.

Train/test split is stratified on `default_payment_next_month` using a 70/30 ratio, preserving class prevalence in both sets. This is crucial because all subsequent estimates (AUC, Brier, cost curves) are interpreted under the observed default rate.

3 Method

Feature engineering. To avoid leakage we apply identical transformations to train and test. We log-transform `LIMIT_BAL` and `AGE` via $\log(1+x)$ and build four blocks of behavioural features: (i) utilisation (monthly bill/limit ratios and simple summaries such as mean, max and last, plus a count of high-utilisation months with utilisation ≥ 0.9); (ii) delay history (number of late months, maximum and last delay, and an indicator of severe delay ≥ 2 months); (iii) repayment behaviour (monthly pay/bill ratios, their mean/min/last and a count of underpayment months with ratio < 0.2); and (iv) debt level and trend (total bills and payments, pay-to-bill and pay-to-limit ratios, and a bill-trend indicator). These blocks summarise intensity of line usage and repayment discipline, i.e. the core drivers for PD decisions.

Multicollinearity. On the full logistic model we compute adjusted GVIF; values range from about 1 to 3.8 and no predictor exceeds $\text{GVIF}_{\text{adj}} > 4$. Multicollinearity is thus moderate but acceptable, so we retain the utilisation block as a joint signal of stress on the credit line.

Models. We estimate logistic regressions

$$\mathrm{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta^\top x_i,$$

with dummy encoding for categorical predictors, and fit three nested specifications: *demo* (demographics and static factors), *core* (demo + key behavioural indicators) and *full* (demo + the full behavioural set). AIC/BIC and McFadden pseudo-R^2 favour the full model, which provides the best fit while remaining interpretable in odds-ratio terms.

Random forest. As a non-parametric benchmark we train a random forest classifier with probability output using `ranger` (500 trees, standard *mtry* and minimum node size). Variable-importance rankings (Gini decrease) offer a complementary view of drivers, while out-of-bag predictions are used to monitor overfitting under slightly more conservative hyperparameter settings.

Calibration and imbalance. We calibrate raw scores s_i (logit or random forest) via Platt scaling, i.e. a second logistic mapping learned on the training set and applied to predicted scores. The full logistic model is already close to identity, whereas the random forest is overconfident and benefits markedly from calibration. In parallel we compare three full logistic variants (`logit_full_unweighted`, `logit_full_weighted`, `logit_full_balanced`) trained on the same features and evaluated on the original test set to assess whether class reweighting or resampling adds value beyond threshold tuning.

Segmentation. Finally, we run k-means on standardised behavioural variables, select $k = 4$ via elbow and silhouette criteria, and assign test clients to the nearest centroid. One cluster (segment 4) stands out as high risk (default rate about 35%) and is later used to illustrate segment-specific PD thresholds and differentiated policies.

4 Implementation

Tools and data flow. The pipeline is implemented in R with a reproducible script structure and relative paths via `here::here`. Clean and processed datasets are stored as `.rds` files and reused across steps. Core packages include `tidyverse` for data handling, `rsample` for splitting, `yardstick` and `pROC` for metrics, `glmnet` and `ranger` for modelling, and `pdp` plus plotting libraries for visualisation. Outputs comprise fitted models, predicted probabilities for train/test, coefficient and GVIF tables, variable-importance rankings, and threshold summaries.

Workflow. Script order mirrors a real risk-modelling process: loading and cleaning, stratified train/test split, feature engineering, logistic and random forest fitting, Platt calibration, discrimination and calibration evaluation, cost-based threshold selection, k-means segmentation, and interpretation via PDP/ICE. The separation between training and test sets is enforced at every stage to prevent leakage and to ensure that performance and thresholds are estimated under realistic conditions.

5 Evaluation

Discrimination. On the held-out test set, performance improves mainly when moving from demographics to behavioural information. The `logit_demo` model reaches ROC AUC ≈ 0.62–0.63 (PR AUC ≈ 0.31). Adding key behavioural features (`logit_core`) raises ROC AUC to ≈ 0.76 (PR AUC ≈ 0.51). The full logistic with calibration (`logit_full_platt`) attains ROC AUC ≈ 0.77–0.78 and PR AUC ≈ 0.53. The calibrated random forest (`rf_full_platt`) matches logit in ROC AUC and slightly improves PR AUC (to ≈ 0.56), but does not materially dominate it. The main gain thus comes from behavioural features rather than from the choice of algorithm.

Calibration. Calibration is assessed via the Brier score and its Murphy decomposition. With default prevalence around 22%, `logit_full` achieves a Brier score of about 0.134 on both train and test, with a small reliability term and stable resolution; Platt scaling leaves PDs essentially unchanged, indicating that the model already produces well-aligned probabilities. For `rf_full`, Platt calibration is more impactful: it corrects strong overconfidence on train and yields test Brier scores around 0.134–0.136, comparable to the logistic but with slightly higher resolution and a small residual reliability term. In practice, both calibrated models can be used as PD engines.

Imbalance variants. For the three full logistic variants (unweighted, weighted, balanced), ROC AUC on the test set is almost identical (0.79). However, the unweighted model attains the best Brier score and the lowest minimum cost under standard FN:FP ratios, while weighted and balanced versions produce worse calibration and implausibly high optimal thresholds. In this setting, keeping the original class distribution and tuning thresholds is preferable.

Threshold selection. For `logit_full_platt`, the Youden index is maximised at a threshold $\tau \approx 0.23$, giving sensitivity ≈ 0.64, specificity ≈ 0.80 and accuracy ≈ 0.77. When asymmetric costs are introduced (e.g. FN:FP = 5:1), the cost-minimising threshold moves down to around $\tau \approx 0.17$, increasing sensitivity at the expense of more false positives. For `rf_full_platt`, cost-optimal thresholds lie in a similar band (roughly 0.16–0.20), confirming that, once calibrated, the choice of model does not radically change the decision frontier.

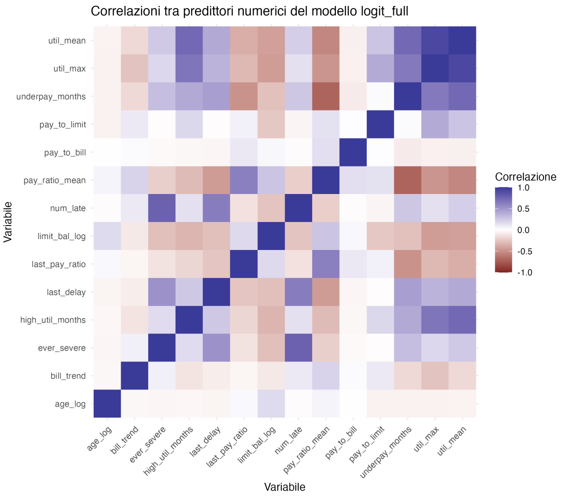
Thresholds by segment. Using the $k = 4$ segmentation fitted on train, we compute segment-specific Youden and cost-optimal thresholds and apply them to test. The high-risk segment 4 (default rate around 35%) exhibits higher optimal cut-offs than the portfolio as a whole, reflecting the fact that medium PDs are already economically "high" in this group.

6 Interpretation

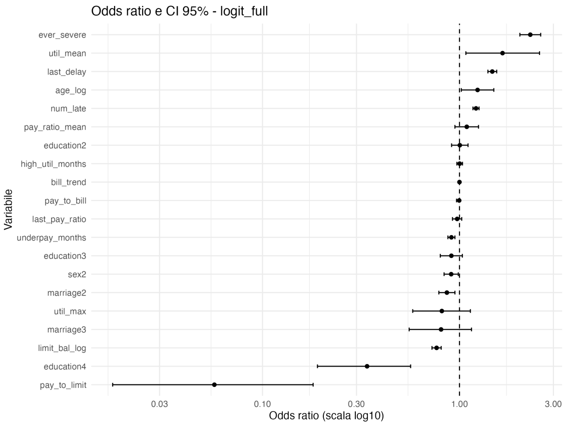
Behavioural drivers. Both the logistic regression and the random forest converge on the same behavioural drivers of PD. Recent delays and any episode of severe arrears sharply increase default odds, with the largest jump in risk occurring when a client moves from zero to one late month. High and persistent utilisation of the credit line signals structural financial stress, rather than temporary liquidity shocks. Even in the absence of formal delinquency, weak repayment discipline (low or unstable repayment ratios, repeated underpayments) and a rising debt trend further raise predicted PD.

PDP and ICE evidence. PDP and ICE analyses confirm essentially monotone relationships: more severe or more recent delays and higher utilisation steadily increase PD, whereas stronger and more regular repayment ratios reduce it. Individual variation around these curves does not overturn the global patterns, which explains why a well-specified, calibrated logistic model already captures most of the actionable signal. The random forest mainly refines the picture by revealing practical thresholds where risk jumps (for example, the first severe delay or the first late month), which are useful for rule-based triage.

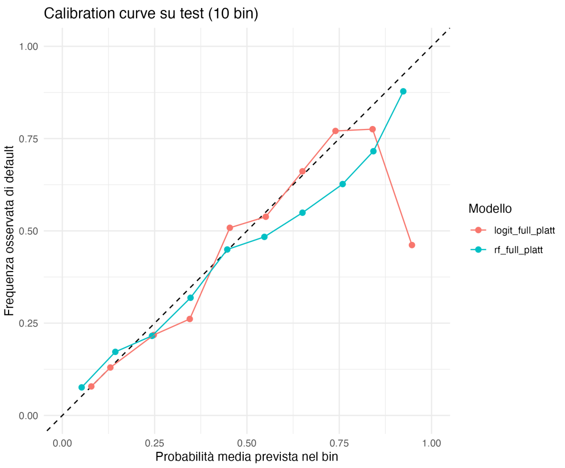
Economic implications and governance. Economically, this parsimonious and interpretable model supports transparent thresholds, stable credit policies and clear communication with supervisors. Linking calibrated one-month PDs to explicit cost ratios and segment-specific rules allows the bank to adjust approval, limit management and pricing to different risk appetites. A simple governance framework—calibration monitoring, challenger models and fairness checks—further favours a well-understood, robust model over alternatives that are slightly more predictive but more opaque and fragile.



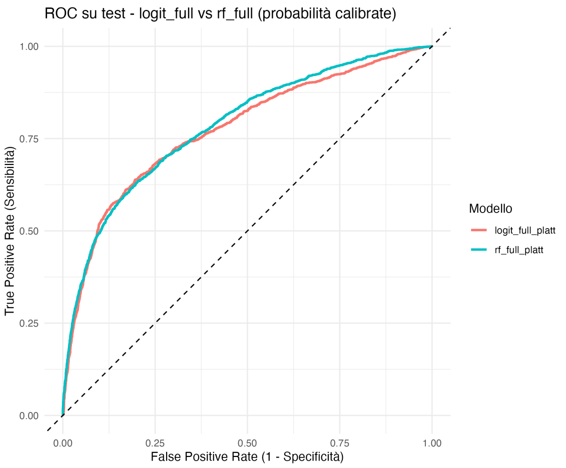
(a) Correlation matrix of key numerical predictors used in the full logistic model.



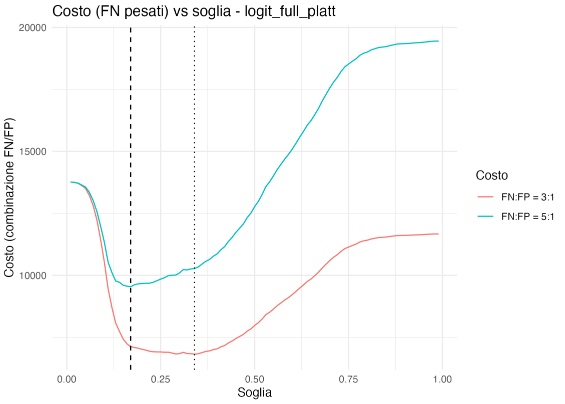
(b) Odds ratios (with 95% confidence intervals) for the full logistic regression model.



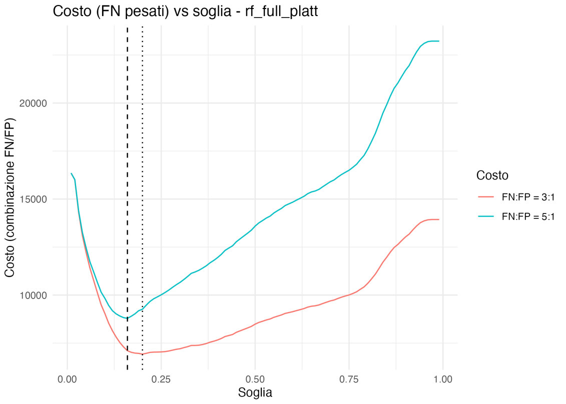
(c) Calibration curve on the test set for the calibrated full logistic model and the calibrated random forest (10 bins).



(d) ROC curves on the test set for the calibrated full logistic model and calibrated random forest.



(e) Expected misclassification cost vs. threshold for the calibrated full logistic model (FN:FP = 3:1 and 5:1).



(f) Expected misclassification cost vs. threshold for the calibrated random forest model (FN:FP = 3:1 and 5:1).