Arnur Abdrakhman

Data Bootcamp Fall 2025 – Final Project

Introduction

This project is a credit default risk tool that develops and evaluates predictive models based on a real world consumer credit dataset.The predictive task is to determine the likelihood of a borrower defaulting on his or her credit obligations, with applications for credit decision policies.

Several models are compared, namely logistic regression, random forest and histogram-based gradient boosting. Although tree-based models have higher predictive accuracy, logistic regression is emphasized because it is easier to interpret and in line with conventional credit risk practice. Along with the usual accuracy measures, threshold-tuning, cost-sensitive decision-making rules, and tradeoff between false positives and false negatives are explored in the project as realistic financial decision-making constraints.

Data Description

The dataset used is the UCI "Default of Credit Card Clients" dataset (Taiwan) which includes 30,000 observations and 23 input features.The target variable shows whether or not a client has defaulted on their credit obligation (1 = default, 0 = non-default). The main features of the data:

- Default rate: 22.12% (mean of target = 0.2212)
- Features include credit limit, demographic variables, payment history indicators, and bill/payment amounts.
- None of the missing values were identified in the key input features.

The data is representative of an average retail credit portfolio with moderate class imbalance and hence is appropriate to assess predictive performance as well as policy-based threshold selection.

Models and Methods

Model Specification

Three classification models were adopted and compared:

1. Logistic Regression

    - Regularized (L1 / L2)
    - Taken to be the primary interpretable model
    - In line with scorecard-based credit risk modeling

2. Random Forest Classifier

    - Ensemble-based non-linear model
    - Used as a performance benchmark

3. HistGradientBoosting Classifier (HGB)

    - Gradient-boosted tree model
    - Obtained the strongest overall predictive performance

All models were trained using a train–test split and evaluated on the same held-out test set.

Evaluation Metrics

The evaluation of performance was based on:

i. ROC–AUC (ranking quality)
ii. Precision, recall, and F1-score of default class
iii. Confusion matrices to study decision consequences

Results and Interpretation

Baseline Model Performance

Test-set ROC–AUC results show a certain ranking by performance:

- Logistic regression: ROC–AUC ≈ 0.716
- Random forest: ROC–AUC ≈ 0.752
- HistGradientBoosting: ROC–AUC ≈ 0.779

This means that the tree based models offer more separation between defaulters and non-defaulters. Nevertheless, increased AUC does not necessarily determine the suitability of a model to be deployed.

Threshold Tuning with a Recall-Focused Policy

It initial analysis considered a recall-based threshold, which was built upon the notion that a defaulter miss (false negative) is typically more expensive than a good borrower rejection.

In the case of logistic regression, a threshold of 75% recall on defaulters gave:

- Threshold ≈ 0.10
- Default recall ≈ 99.6%
- Default precision ≈ 22.2%
- A very large increase in false positives (> 5,700 good borrowers flagged as risky)

This policy highly emphasizes risk avoidance and results in aggressive rejection behavior, which is an example of the cost of being extremely conservative in credit screening.

Threshold Tuning with Maximum F1-Score

Thresholds were then chosen to maximize the F1-score of the default class in order to balance the trade off between precision and recall.

Results under max-F1 thresholds:

- Logistic Regression
    - Threshold ≈ 0.60
    - Default precision ≈ 0.564
    - Default recall ≈ 0.459
    - Default F1 ≈ 0.506
- Random Forest
    - Threshold ≈ 0.31
    - Default F1 ≈ 0.525
- HistGradientBoosting
    - Threshold ≈ 0.30
    - Default precision ≈ 0.535

- Default recall ≈ 0.555
- Default F1 ≈ 0.545

Using this balanced criterion, HistGradientBoosting provides the best overall performance in terms of classification, fewer false negatives and a more stable tradeoff than logistic regression.

Cost-Sensitive Analysis

Recognizing that financial institutions vary in their risk tolerance, expected-cost thresholding was investigated:

$$\text{Expected Cost} = C_{FN} \cdot FN + C_{FP} \cdot FP$$

Three cost ratios were studied, including 5:1, 10:1 and 20:1, where missing a default is assumed to be 5-20 times more expensive than turning away a good borrower.

Key findings:

- As the cost of false negatives increases, optimum thresholds become smaller and recall takes precedence.
- Logistic regression reduces to extremely conservative policies at increased cost ratios (threshold ≈ 0.04 at 10:1 and 20:1), which amounts to rejecting most applicants.
- HistGradientBoosting is always cheaper across all cost ratios.

At a 20:1 cost ratio, the HGB model selected a threshold ≈ 0.06, giving:

- FN = 28
- FP = 5,150
- Almost total capture of defaulters, at the cost of a good deal of lost business.

This points to the fact that the best policy is one that is based on institutional preferences, rather than statistical performance.

Model Interpretation

Interpretability analysis was conducted only for the logistic regression model, which is in line with most credit risk practice. The most significant coefficients in absolute value are X6, X12, X19, and X18, which reveal that there is a strong relationship between recent payment behavior and default risk.

Permutation importance analysis confirms that these same features materially affect predictive performance, reinforcing their relevance. Although tree-based models are more accurate than

logistic regression, their absence of clear coefficients restricts the use of these models to direct regulatory or customer explanations.

Conclusion and Next Steps

This project demonstrates that model selection and threshold choice are policy decisions, not purely technical ones. HistGradientBoosting offers the highest predictive accuracy and cost-effectiveness especially when there are asymmetric loss assumptions. Although less accurate, logistic regression provides transparency and stability that are still at the core of credit risk governance.

Rather than prescribing a single “best” model, the analysis suggests:

i. Logistic regression may serve as a baseline or regulatory scorecard
ii. Models based on trees can be either challenger models or internal risk engines.
iii. Thresholds should be chosen based on institutional risk appetite, capital constraints, and market strategy

Next Steps

Potential extensions include:

i. Probability calibration and stability analysis
ii. Time-based validation to assess performance drift
iii. Explicit profit or capital-based objective functions
iv. Fairness and segmentation examination across borrower groups

In general, the findings highlight that successful credit risk modeling should provide a balance between statistical performance, financial goals and governance constraints, leaving room for bank-specific preferences and judgment.