Project 2: Credit Risk and Statistical Learning

Deadline: December 22, 2025

Contact: Andrea Ruglioni (andrea.ruglioni@epfl.ch)

General instructions. You may work in groups of up to three students (smaller groups are also allowed). Your submission must contain:

1. A single **PDF report** including a clear description of your methodology, the results of your experiments (tables and figures), and your *commented code* in an appendix.
2. The corresponding **code files** in Python (.ipynb or .py).

The PDF should be fully self-contained: all results, plots, and tables must be visible in the report without executing the code. When the provided scripts are executed, they should exactly reproduce the results shown in your PDF. Please list all group members on the title page of your report.

Grading. This project is worth 10% of the final grade. You will be evaluated based on the correctness of your implementation, the clarity and completeness of your explanations, and the quality of your analysis and interpretation of results.

Project description. This project focuses on the quantitative modeling of credit risk for retail loans. You will work with *synthetic* data where each loan applicant is described by three characteristics: age, income, and employment type. Your goal is to build and evaluate statistical learning models that predict repayment probabilities and to assess the profitability and risk of different lending strategies.

1. **Feature generation.** Simulate $m+n$ feature vectors $x^i = (x_1^i, x_2^i, x_3^i) \in \mathbb{R}^3$, $i = 1, \dots, m+n$, with $m = 20000$ and $n = 10000$, where:
   - x_1^i: age, uniformly distributed on $[18, 80]$,
   - x_2^i: monthly income (in thousands of CHF), uniformly distributed on $[1, 15]$,
   - x_3^i: employment status in $\{0, 1\}$ with $\mathbb{P}(x_3^i = 1) = 0.1$ (self-employed) and $\mathbb{P}(x_3^i = 0) = 0.9$ (salaried).

   Assume independence across the three coordinates.

   a) Compute empirical means and standard deviations of each feature based on the first m samples.

   b) Suggest other variables that would realistically be relevant in credit scoring (beyond age, income, and employment type).

2. **Default model and data generation.** Let $(\xi^i)_{i=1}^{m+n}$ be i.i.d. Unif$(0, 1)$. Define the sigmoid function as $\psi(z) = 1/(1 + e^{-z})$. Consider two repayment probability functions $p_1, p_2 : \mathbb{R}^3 \to (0, 1)$:

$$p_1(x) = \psi(13.3 - 0.33x_1 + 3.5x_2 - 3x_3),$$
$$p_2(x) = \psi\big(5 - 10\big[\mathbf{1}_{(-\infty,25)}(x_1) + \mathbf{1}_{(75,\infty)}(x_1)\big] + 1.1x_2 - x_3\big).$$

Construct two datasets (x^i, y_1^i) and (x^i, y_2^i), $i = 1, \ldots, m+n$, by setting

$$y_s^i = \begin{cases} 1, & \text{if } \xi^i \le p_s(x^i), \\ 0, & \text{otherwise,} \end{cases} \qquad s = 1, 2.$$

Interpretation: $y_s^i = 1$ means borrower i repays, $y_s^i = 0$ means default. For each dataset $s = 1, 2$:

a) Fit a logistic regression model $\hat{p}_s^{\log} : \mathbb{R}^3 \to \mathbb{R}$ on the training data (x^i, y_s^i), $i = 1, \ldots, m$. Report cross-entropy loss on both training and test sets ($i = m+1, \ldots, m+n$). You can use the Python function `sklearn.linear_model.LogisticRegression`.

b) Standardize features by dividing each coordinate by its empirical standard deviation from the training set. Fit a Support Vector Machine (SVM) classifier with Gaussian kernel

$$k(x, x') = \exp\left(-\frac{1}{10}\|x - x'\|_2^2\right),$$

using hinge loss and regularization parameter $\lambda = \frac{5}{2m}$. In `sklearn.svm.SVC`, this corresponds to $C = 0.2$. Use the option `probability=True` to obtain probability estimates $\hat{p}_s^{\text{svm}}$ (Platt scaling[1]). Evaluate cross-entropy loss of $\hat{p}_s^{\text{svm}}$ on both training and test sets.

c) Plot the ROC curves, i.e., graphs of the True Positive Rate (TPR) versus the False Positive Rate (FPR), using the test data. Additionally, compute the corresponding Area Under the ROC Curve (AUC) for $\hat{p}_s^{\log}$ and $\hat{p}_s^{\text{svm}}$.

3. **Lending strategies.** Focus now on dataset 2, i.e., (x^i, y_2^i). Assume each loan is of CHF 1000 and that repayment is all-or-nothing (full repayment with interest, or total default). Consider the following lending policies applied to the test set ($i = m+1, \ldots, m+n$):

(i) Lend to everyone at 5.5% interest.

(ii) Lend selectively at 1% interest, but only to applicants with $\hat{p}_2^{\log}(x^i) \ge 95\%$.

(iii) Same as (ii), but selection based on $\hat{p}_2^{\text{svm}}$.

To evaluate performance, simulate 50000 scenarios of repayment outcomes: for each test applicant i, draw independent $\xi^{i,k} \sim \text{Unif}(0,1)$ and set

$$D_{i,k} = \mathbf{1}\{\xi^{i,k} \le p_2(x^{m+i})\}, \qquad i = 1, \ldots, n,\ k = 1, \ldots, 50000.$$

Here $D_{i,k} = 1$ indicates repayment in scenario k. For each strategy (i)–(iii):

a) Plot the histogram of profits & losses across the 50000 scenarios and compute the expected profit & loss.

b) Estimate the 95%-VaR and 95%-ES of the profit & loss distribution.

[1] `https://home.cs.colorado.edu/~mozer/Teaching/syllabi/6622/papers/Platt1999.pdf`