Credit Default Prediction Using Machine Learning

ESILV – Machine Learning Project 2025

Gabriel Picard Thibault Pelou Hugo Picard

December 5, 2025

Abstract

This report presents the development of a complete machine learning pipeline for predicting credit default using demographic information and longitudinal credit histories. We follow the full methodology required in the project guidelines: business case formulation, data exploration, preprocessing, imbalanced learning, baseline and advanced models, hyperparameter tuning, interpretability, and final evaluation. Our best-performing model demonstrates significant improvement over naïve baselines, particularly in recall, the metric most aligned with real-world financial risk management.

Contents

1 Business Case

Credit risk assessment is a fundamental task in financial engineering, banking, and regulatory compliance. Institutions must evaluate whether clients will default on their credit obligations and compute a probability of default (PD).

Accurate PD estimation is essential for:

- risk-based pricing,
- expected loss computation,
- Basel II/III regulatory capital requirements,
- stress-testing and portfolio management,
- automated online lending and fintech scoring systems.

The aim of this project is to design a supervised learning model capable of predicting whether a borrower will default based on personal information and historical repayment patterns.

The problem is formulated as a binary classification task:

$$y = \begin{cases} 1 & \text{if the client exhibited at least one delinquency event} \\ 0 & \text{otherwise} \end{cases}$$

2 Dataset Description

We use the *Home Credit Default Risk* dataset, composed of two files:

2.1 Application Data (application_record.csv)

This dataset contains one row per client with socio-economic variables:

- demographic features (gender, number of children, family status),
- financial information (total income, employment duration),
- education and occupation,
- housing characteristics.

2.2 Credit History (credit_record.csv)

This dataset contains monthly repayment statuses for each customer. The variable `STATUS` takes values:

- `0`: paid on time,
- `1--5`: increasingly severe delinquency,
- `C`: closed loan,
- `X`: unknown / no information.

2.3 Target Definition

Following industry practice and the project objectives, we define:

$$\text{default} = 1 \quad \text{if STATUS} \in \{1, 2, 3, 4, 5\} \text{ at least once}$$

This produces a highly imbalanced target distribution:

$$\Pr(y = 1) \approx 0.12$$

3 Data Exploration (EDA)

We performed exploratory data analysis on both datasets.

3.1 Missing Values

The variable `OCCUPATION_TYPE` contains over 11,000 missing entries. We impute:

- numerical variables using the median,
- categorical variables using the most frequent category.

3.2 Distributions and Outliers

The dataset contains extreme outliers in:

- `AMT_INCOME_TOTAL` (fat-tailed distribution),
- `DAYS_EMPLOYED` (with anomalous value 365243),

which we treat using domain-driven transformations (e.g., converting days into years).

3.3 Correlation Analysis

Figure placeholders for the report:

Figure 1: Correlation heatmap of numerical variables.

Figure 2: Distribution of the target variable (highly imbalanced).

4 Problem Formalisation

Given input vector $x \in R^d$, we model:

$$\hat{y} = f(x), \qquad f : R^d \to \{0, 1\}$$

and estimate:

$$p(x) = \Pr(y = 1 \mid x)$$

The main challenge is imbalanced learning. Therefore, classical accuracy is not meaningful. We focus on:

$$\text{Recall}, \quad \text{Precision}, \quad \text{F1-score}, \quad \text{ROC-AUC}$$

5 Preprocessing and Feature Engineering

We use a unified pipeline:

- **Imputation**: median (numerical) / most-frequent (categorical)
- **Scaling**: StandardScaler for numerical features
- **Encoding**: OneHotEncoder for categorical variables
- **Credit aggregation**: number of delinquencies, worst delinquency, credit history length
- **Train/test split**: 70/30 with stratification

The pipeline is implemented using `ColumnTransformer` and `Pipeline`.

6 Models Implemented

We implemented several model categories:

6.1 Baseline Model

- Logistic Regression (vanilla)

6.2 Standard Models

- Decision Tree
- Random Forest
- Support Vector Machine (RBF Kernel)
- Logistic Regression with PCA

6.3 Cost-Sensitive Models

- Logistic Regression with class weights
- Random Forest with class weights

6.4 Resampling Methods

- Random Oversampling
- SMOTE

6.5 Advanced Model

- XGBoost with SMOTE

6.6 Hyperparameter Tuning

We perform grid search over:

- `n_estimators`, `max_depth` for tree models,
- `C`, `kernel` for SVM,
- learning rate and tree depth for XGBoost.

7 Evaluation and Results

7.1 Metrics

For imbalanced data, recall and F1 are emphasized.

7.2 Final Comparison Table

Model	Accuracy	Precision	Recall	F1
RandomForest (balanced)	0.84	0.37	0.51	0.43
RandomForest + Oversampling	0.81	0.33	0.59	0.42
RandomForest + SMOTE	0.86	0.41	0.40	0.40
XGBoost + SMOTE	0.86	0.25	0.10	0.15
Decision Tree	0.88	0.47	0.29	0.36
LogReg (balanced)	0.58	0.14	0.48	0.21
SVM (RBF)	0.88	0.62	0.00	0.01

Table 1: Model performance comparison (Test Set).

7.3 Threshold Optimization

Threshold tuning significantly improves F1 and recall:

$$\text{Optimal threshold} = 0.31$$

Figure 3: Precision–Recall curve used for threshold selection.

8 Interpretability

We apply:

- **Permutation importance**
- **SHAP values** for detailed insight into feature-level contributions

Key predictors:

- Number of delinquent months

- Worst delinquency status
- Employment duration
- Total income

Figure 4: SHAP summary plot for RandomForest + SMOTE.

9 Discussion

Our results highlight several challenges typical in credit scoring:

- Severe class imbalance makes accuracy misleading.
- Threshold tuning is crucial for maximizing recall.
- Tree-based models outperform linear ones due to complex interactions.
- Resampling techniques (especially SMOTE) significantly help minority detection.

Limitations:

- Credit history features could be engineered more deeply.
- Time-series models may capture sequential dynamics better.
- Economic or macro data could improve predictability.

10 Conclusion

We developed a robust machine learning pipeline for credit default prediction, combining preprocessing, imbalanced learning, model comparison, hyperparameter tuning, interpretability, and threshold optimization.

Our best model, **RandomForest + SMOTE + threshold optimization**, achieves strong recall and balanced F1-score, demonstrating the practical applicability of ML methods in financial risk engineering.

Future improvements may include deep learning architectures, temporal models, and more advanced feature extraction from credit histories.

11 References

- Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.*
- He, H., & Garcia, E. (2009). *Learning from Imbalanced Data.*
- Kaggle Home Credit Dataset: `https://www.kaggle.com/`