Machine Learning Project

Credit Risk Classification Using Machine Learning Models

Jack Liu
ESILV – De Vinci Higher Education

December 2025

Contents

1 Business Scope and Problem Definition

1.1 Context and Business Case

Describe the financial context of credit risk management and explain why predicting credit default is a crucial task for financial institutions.

1.2 Project Objectives

Clearly define the objective of the project, the target variable, and the type of machine learning problem addressed (binary classification).

2 Dataset Description

2.1 Dataset Source

Present the dataset source, its origin, and its relevance to the business problem.

2.2 Variables Description

Describe numerical and categorical variables and explain their economic meaning when relevant.

3 Exploratory Data Analysis

3.1 General Statistics

Number of observations, number of features, missing values, and data types.

3.2 Data Visualization

Histograms, boxplots, and correlation heatmaps.

3.3 Class Imbalance Analysis

Analyze the distribution of the target variable and discuss imbalance issues.

4 Data Preprocessing

4.1 Data Cleaning

Handling missing values, duplicates, and outliers.

4.2 Encoding and Scaling

Explain encoding methods for categorical variables and feature scaling techniques.

4.3 Train-Test Split

Describe the train-test split strategy and justify the choice.

5 Baseline Model

5.1 Logistic Regression

Presentation of the baseline Logistic Regression model.

5.2 Baseline Results

Evaluation using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix.

6 Handling Class Imbalance

6.1 Imbalance Problem

Explain why class imbalance is an issue in this dataset.

6.2 Resampling Techniques

Describe SMOTE, Random Over-Sampling, and other techniques used.

6.3 Impact on Model Performance

Compare model performance before and after resampling.

7 Hyperparameter Optimization

7.1 Cross-Validation Strategy

Explain the use of Stratified K-Fold cross-validation.

7.2 Grid Search

Describe the hyperparameter tuning process and selected parameters.

8 Advanced Machine Learning Models

8.1 Decision Tree

Model description and results.

8.2 Random Forest

Model description, advantages, and results.

8.3 Gradient Boosting

Explain boosting principles and present results.

8.4 Support Vector Machine

Kernel choice, tuning, and performance.

8.5 XGBoost or LightGBM

Justification of the model choice and reference to scientific literature.

9 Dimensionality Reduction

9.1 Principal Component Analysis

Explain PCA methodology and variance explained.

9.2 Impact on Performance

Compare models with and without PCA.

10 Ensemble Learning

10.1 Voting Classifier

Explain ensemble strategy and combined results.

10.2 Performance Comparison

Compare ensemble models with individual models.

11 Model Comparison

11.1 Quantitative Comparison

Present a comparison table of all models.

11.2 Qualitative Analysis

Discuss interpretability, robustness, and business relevance.

12 Limitations

Discuss dataset limitations, model assumptions, and potential biases.

13 Conclusion and Perspectives

Summarize the main findings and propose future improvements.

14 References

References

[1] Scikit-learn Documentation, `https://scikit-learn.org`

[2] Chen, T. and Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD.