Project 3 Proposal. Ensemble Learning for Credit Card Default Prediction

Laura González and Eva Martín

December 21, 2025

Project Specification

This project addresses a supervised binary classification problem in the context of credit risk modeling, focusing on the prediction of credit card default. We consider the *Default of Credit Card Clients* dataset, originally introduced by Yeh and Lien (2009), which contains 30,000 real-world credit card records from Taiwan, described by demographic variables, credit limits, historical repayment status, bill amounts, and past payments over a six-month period.

The dataset represents a challenging and realistic learning problem due to its moderate class imbalance, heterogeneous feature types, and the presence of nonlinear interactions among financial variables. These characteristics make it well suited for the application of advanced ensemble learning methods, which are known to perform strongly on tabular data with complex dependency structures.

The goal of the project is to study and compare different ensemble paradigms from Part III of the course, namely bagging-based methods (Random Forests), boosting-based methods (Gradient Boosting and Extreme Gradient Boosting), and stacking strategies that combine heterogeneous base learners. Rather than focusing solely on predictive performance, the project aims to analyze how these ensemble techniques differ in terms of bias–variance trade-offs, model diversity, stability, and interpretability in a credit risk setting.

The project will be carried out using standard CPU-based computational resources and established machine learning libraries. All experiments will follow a rigorous and reproducible evaluation protocol, enabling a fair comparison across ensemble methods.

Justification and Learning Goals

Credit risk prediction is a classical problem in applied machine learning, yet it remains challenging due to the nonlinear relationships, heterogeneous variables, and evolving patterns present in real financial data. The Default of Credit Card Clients dataset has been widely used as a benchmark in the literature.

While early studies focused on individual classifiers such as logistic regression, neural networks, or decision trees, more recent work has shown that ensemble methods can provide substantial improvements in predictive performance. In particular, bagging, boosting, and hybrid ensemble strategies have been reported to outperform single models in credit default prediction, highlighting the importance of model diversity and error reduction through aggregation.

The primary motivation of this project is to deepen our understanding of advanced ensemble learning techniques covered in Part III of the course and to analyze their behavior in a realistic application domain. This project focuses on a structured comparison of different ensemble approaches, aiming to understand how and why their performance differs in practice. In particular, we are interested in the trade-offs between predictive accuracy, model stability, computational cost and interpretability.

Methodology and Theoretical Foundations

The modelling strategy will involve:

- A bagging-based ensemble using Random Forests, serving as a strong baseline to study variance reduction through aggregation of decision trees.
- Boosting-based models, including Gradient Boosting and Extreme Gradient Boosting (XGBoost), to analyze bias reduction through sequential additive modeling.

- A stacking approach combining heterogeneous base learners, with a meta-model trained to exploit complementary strengths across models.

All models will be trained and evaluated under a common experimental protocol to ensure a fair comparison. Hyperparameters will be selected using cross-validation, and performance will be assessed using metrics appropriate for imbalanced classification problems, such as ROC-AUC and precision-recall measures.

Fundamental References

1. AML Course Material: Part III Theory Notes and Lab Sessions (Racó).
2. I.-C. Yeh and C.-H. Lien (2009). *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients.* Expert Systems with Applications, 36(2), 2473–2480.
3. L. Breiman (2001). *Random Forests.* Machine Learning, 45, 5-32.
4. J. H. Friedman (2001). *Greedy Function Approximation: A Gradient Boosting Machine.* Annals of Statistics, 29(5).
5. T. Chen and C. Guestrin (2016). *XGBoost: A Scalable Tree Boosting System.* Proceedings of KDD.
6. S. R. Islam, W. Eberle, and S. K. Ghafoor (2018). *Credit Default Mining Using Combined Machine Learning and Heuristic Approach.* arXiv preprint.

you should make the models "hybrid ensemble strategies" more precise - the study of a very recent model would make the project much more interesting: https://arxiv.org/abs/2508.10053 - before start, make sure you detail the specific quality measures (be sure to include at least a handful).