MACHINE LEARNING CASE STUDY

Credit Risk Scoring with Ensemble Methods

Roger Estrada · Machine Learning in Business · AI Strategy · 2026

Random Forest ranked loan applicants best among five models (AUC 0.776), and every ensemble clearly beat a single decision tree (AUC 0.606). Choosing the model on AUC rather than accuracy matters because the approval threshold will be set by the cost of each type of error.

Business problem

Underwriting needs a defensible, repeatable way to score applicants. The errors are asymmetric: approving a bad loan loses principal, while declining a good one loses interest income. The real deliverable is the model and threshold that fit that asymmetry, along with an explanation a credit committee can accept.

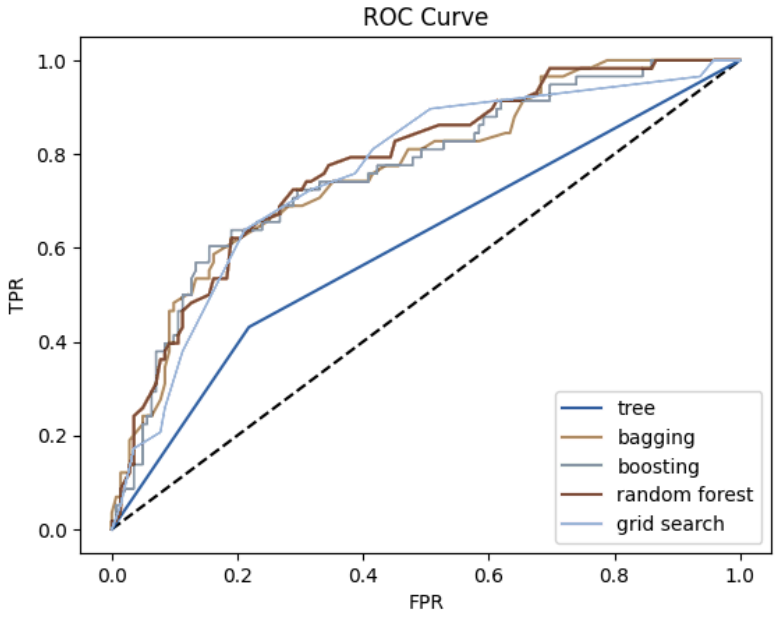
Approach and results

Data. The German Credit benchmark: 1,000 loan applicants described by 16 attributes covering checking and savings balances, loan amount and duration, credit history, purpose, employment, age, and housing. The outcome is default or no default. The data was split 80/20 with a fixed random seed.

Method. Five models were built and compared on the same test set: a single decision tree, a tuned decision tree (GridSearchCV with 10-fold cross-validation over depth, leaf count, and split criterion), Bagging, AdaBoost, and Random Forest, each ensemble with 100 trees. AUC was the primary metric.

Model	Test accuracy	Test AUC
Single decision tree	0.680	0.606
Tuned decision tree (depth 10, 17 leaves)	0.740	0.759
AdaBoost	0.765	0.763
Bagging	0.770	0.764
Random Forest	0.755	0.776

The single tree fit its training data perfectly (accuracy 1.000) and scored 0.68 on new applicants, confirmed by five-fold cross-validation at 0.68. Every ensemble lifted AUC by 0.16 to 0.17 over the single tree. The accuracy leader (Bagging, 0.770) is not the AUC leader (Random Forest, 0.776), which is the reason the selection metric matters.



What drives the score. Random Forest's most influential inputs were loan amount (0.142), loan duration in months (0.109), applicant age (0.105), and checking-account status (0.067).

Enterprise application

The workflow transfers to consumer and business lending, credit limit assignment, insurance underwriting, and supplier risk. Any extend-or-decline decision that repeats hundreds or thousands of times a year benefits from consistent scoring.

Automation handles the clear approvals and declines. Borderline scores route to a human underwriter. That shortens decision time while accountability stays with qualified people.

Governance and limits

- **Model risk.** Credit models need documented development, independent validation, and ongoing performance monitoring.
- **Fair lending.** Inputs such as age require review, and adverse-action reasons must be explainable.
- **Explainability.** Ensembles raise accuracy but reduce transparency. A production design pairs them with explanation methods or keeps an interpretable challenger model.
- **Scale.** The benchmark has 1,000 applicants. It demonstrates method and metric discipline and is not a production credit model.