Credit Risk Prediction Using Machine Learning

Predicting whether a loan applicant will default using historical financial and demographic data.

By: Abdelmoutaleb Benlyazid

2025/2026

This project explores credit risk prediction, a key problem in the financial sector where lenders must decide whether a loan applicant is likely to default. Building accurate models for this task is critical for reducing financial losses while ensuring fair access to credit.

The work was carried out as part of my summer learning, with the goal of going beyond theory and focusing on the practical, hands-on aspects of data science. I wanted to replicate the full end-to-end workflow that data professionals face in real projects: cleaning raw data, handling imbalances, training multiple machine learning models, tuning hyperparameters, and evaluating results with appropriate metrics.

By approaching the project this way, I aimed to prepare myself for future professional challenges and strengthen my readiness for real-world applications in business analytics and machine learning.

Introduction

Machine Learning (ML) has become one of the most powerful tools for solving complex, data-driven problems across industries. By learning patterns from historical data, ML models can make accurate predictions and support decision-making in contexts where traditional rule-based approaches fall short.

One domain where ML has had a significant impact is in credit risk assessment. Financial institutions must evaluate whether a loan applicant is likely to repay or default. Traditional credit scoring methods often rely on fixed statistical models and rigid thresholds, which may not capture the complexity of customer behavior. With the growth of digital data, ML offers a more flexible and accurate approach to this problem by combining multiple features, handling nonlinear relationships, and adapting to imbalanced datasets.

In this project, the problem is defined as predicting the likelihood of loan default given applicant and loan-related information. The challenge lies in the fact that default cases are relatively rare compared to non-defaults, creating an imbalanced dataset that can mislead standard models.

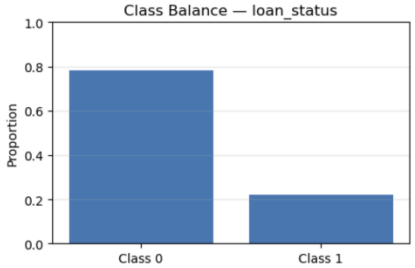
The solution proposed is to design an end-to-end machine learning pipeline that addresses these challenges:

- Cleaning and preparing the data (handling duplicates, missing values, categorical/numerical features).
- Applying oversampling techniques to balance the dataset.
- Training and comparing multiple classifiers (Logistic Regression, Random Forest, XGBoost).
- Using advanced evaluation metrics such as Precision-Recall AUC to better reflect performance in imbalanced settings.
- Exporting the final optimized pipeline for deployment.

This approach provides a practical, real-world workflow that demonstrates how ML can be applied to credit risk prediction in a robust and reliable way.

Data and Methods

The dataset used in this project represents credit applicants and their loan status (default or not). Each record contains a mixture of numerical variables (such as income, loan amount, or duration) and categorical variables (such as marital status or housing situation). The target variable is binary (default = 1, non-default = 0). An important characteristic of the data is its class imbalance: the majority of applicants do not default, while only a smaller fraction represent defaults. This imbalance poses a challenge for predictive modeling and requires careful handling.



ETL and Data Cleaning:

Before modeling, the dataset was cleaned to ensure quality and consistency. The following steps were performed:

- **Duplicate removal** to eliminate repeated entries.
- **Missing value handling**, using median imputation for numerical variables and mode imputation for categorical variables.
- **Consistency checks** to ensure valid ranges and categories.

Preprocessing and Sampling:

To prepare the features for machine learning:

- **Numerical variables** were scaled to a standard range.
- **Categorical variables** were transformed using one-hot encoding.
- **Class imbalance** was addressed by applying **RandomOverSampler** inside the pipeline, ensuring balanced training data while preventing information leakage into validation/test sets.

Modeling Strategy:

Three different models were selected to provide a balance between interpretability and predictive power:

1. **Logistic Regression (LR):** a simple and interpretable baseline model.
2. **Random Forest (RF):** a tree-based ensemble method that handles nonlinearities and feature interactions well.
3. **XGBoost (XGB):** a gradient boosting model known for strong performance on tabular datasets.

Each model was wrapped in a pipeline including preprocessing, sampling, and the classifier. RandomizedSearchCV with cross-validation was used for hyperparameter tuning, with PR-AUC as the primary metric due to the imbalanced nature of the problem.

Results and Evaluation

To evaluate the performance of the models, several metrics were considered. Since the dataset is imbalanced (fewer defaults than non-defaults), accuracy alone would not be informative. Instead, the focus was placed on:

- **Precision-Recall AUC (PR-AUC):** a robust measure for imbalanced datasets that reflects the trade-off between precision (how many predicted defaults were correct) and recall (how many actual defaults were captured).
- **ROC-AUC:** used as a complementary metric to measure the ability of the model to discriminate between the two classes.

- **F1 Score:** the harmonic mean of precision and recall, used for threshold optimization.
- **Confusion Matrix:** provided a clear picture of true positives, false positives, true negatives, and false negatives.

Model Comparison:

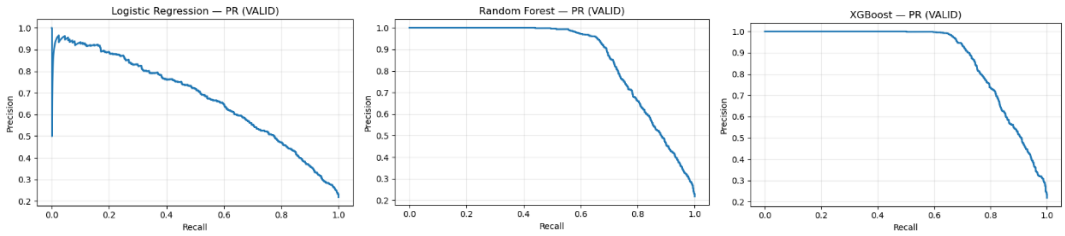
- **Logistic Regression** performed as a solid baseline, offering interpretability but
- moderate predictive power.
- **Random Forest** improved performance by capturing nonlinear relationships and feature interactions.
- **XGBoost** delivered the best overall results, achieving the highest PR-AUC and F1 score across validation sets.

Threshold Optimization:

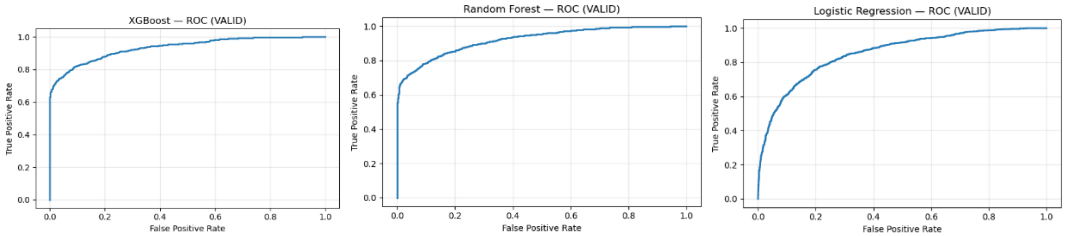
Instead of using the default probability cutoff of 0.5, thresholds were optimized based on F1 scores. This allowed the models, particularly XGBoost, to strike a better balance between capturing defaults (recall) and avoiding false alarms (precision).

Visualization:

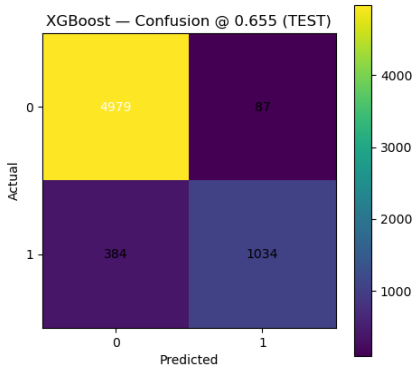
- **Precision-Recall Curves** illustrated the advantage of XGBoost over the other models.



- **ROC Curves** confirmed that tree-based models provided better discrimination.



- **Confusion Matrices** showed the trade-offs in classification, highlighting how oversampling and threshold tuning reduced false negatives (missed defaults).



Overall, the XGBoost model with oversampling and tuned hyperparameters emerged as the most effective solution, providing both strong predictive accuracy and a balanced trade-off between precision and recall.

Conclusion

This project demonstrated how machine learning can be applied to the problem of credit risk prediction, building a complete pipeline from data cleaning to deployment. Through the use of preprocessing, oversampling, and model comparison, it became clear that tree-based models, particularly XGBoost offered the best performance for this dataset, achieving a strong balance between precision and recall as measured by PR-AUC and F1 score.

Beyond the technical results, this work was also a personal learning journey. I gained hands-on experience in:

- Structuring a full ML pipeline from start to finish,
- Handling imbalanced datasets using oversampling,
- Choosing and justifying the right evaluation metrics (PR-AUC over accuracy),
- Applying hyperparameter tuning with RandomizedSearchCV,
- Optimizing thresholds for better business-aligned performance,
- And most importantly, understanding the balance between simplicity (Logistic Regression) and performance (XGBoost) in real-world problems.

This process gave me more confidence in translating machine learning theory into practical solutions, preparing me for professional challenges where end-to-end workflows and clear decision-making are essential.

Project Repository : https://github.com/ABDELMOUTALEB7/credit-risk-model