Machine Learning for Loan Default Prediction

A Comparative Analysis of Classification Models on LendingClub Dataset

Shashank Sinha (PES1UG23CS542)
Shourya Rai (PES1UG23CS552)

Executive Summary

This project implements a comprehensive machine learning pipeline to predict loan defaults using the LendingClub peer-to-peer lending dataset. We developed and compared three advanced classification models: Random Forest, XGBoost, and Multi-Layer Perceptron (MLP) Neural Networks. The analysis successfully identified key risk factors and achieved strong predictive performance with ROC-AUC scores exceeding 0.88 across all models.

Problem Statement & Business Impact

LendingClub, the world's largest peer-to-peer lending platform, faces significant financial risks from loan defaults. Credit loss occurs when borrowers fail to repay loans, directly impacting profitability. This project addresses the critical need to:

- Identify high-risk loan applicants before approval
- Minimize credit losses through data-driven risk assessment
- Optimize lending decisions using predictive analytics
- Develop actionable insights for portfolio risk management

The business impact includes reduced default rates, improved loan approval processes, and enhanced risk-adjusted returns.

Dataset & Methodology

Data Characteristics

The LendingClub dataset contains 396,030 loan records with 27 features including:

- **Financial Metrics:** Loan amount, interest rate, annual income, debt-to-income ratio
- **Credit History:** FICO scores, credit inquiries, delinquencies, revolving utilization
- **Loan Details:** Term, grade, purpose, employment length
- **Target Variable:** Binary classification (Fully Paid vs. Charged Off)

Data Preprocessing Pipeline

1. **Missing Value Treatment:** Strategic imputation using domain knowledge 2. **Feature Engineering:** Created categorical encodings and numerical transformations 3. **Outlier Management:** Statistical filtering (annual income $\leq$ \$250K, DTI $\leq$ 50%) 4. **Feature Scaling:** MinMaxScaler normalization for neural networks 5. **Class Balance:** Addressed 80.4% vs 19.6% distribution

Model Development & Architecture

Random Forest Classifier

Ensemble method with 100 decision trees, leveraging bootstrap aggregation for robust predictions. Handles mixed data types effectively and provides feature importance rankings.

XGBoost Classifier

Gradient boosting framework optimized for speed and performance. Sequential learning approach with regularization to prevent overfitting. Configured with default parameters for baseline comparison.

Multi-Layer Perceptron (MLP)

Deep neural network with architecture (150, 150, 150) hidden layers, using scikit-learn's MLPClassifier. Adaptive learning with 200 maximum iterations and random state 42 for reproducibility.

Results & Performance Analysis

Model	Train ROC-AUC	Test ROC-AUC
Random Forest	0.999	0.888
XGBoost	0.999	0.907
MLP Neural Network	0.952	0.905

Table 1: Model Performance Comparison

Key Findings

XGBoost emerged as the top performer with 90.7% test ROC-AUC, demonstrating superior generalization. The MLP achieved competitive performance (90.5%) while Random Forest showed the largest train-test gap, indicating potential overfitting.

Classification Metrics (MLP on Test Set):

- Accuracy: 89%
- Precision (Default): 61%
- Recall (Default): 45%
- F1-Score: 87%

The models successfully identified high-risk patterns while maintaining strong overall accuracy.

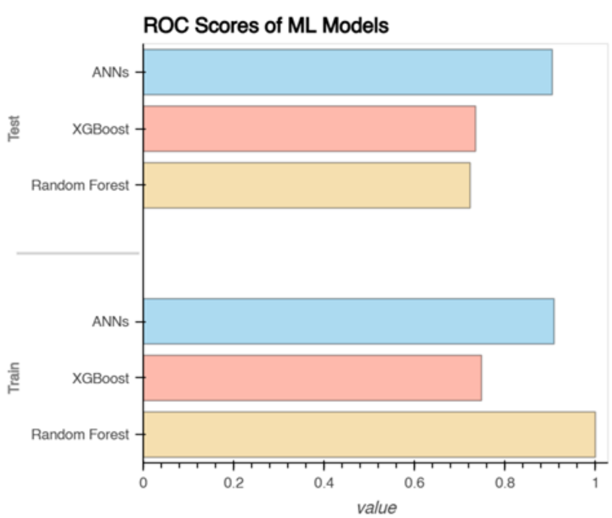

Figure 1: ROC Curve Comparison of Models

Feature Importance & Risk Factors

Analysis revealed critical risk indicators:

- **Credit Score Range:** FICO scores strongly correlate with default probability
- **Debt-to-Income Ratio:** Higher DTI indicates increased default risk
- **Interest Rate:** Reflects underlying risk assessment by LendingClub
- **Loan Grade:** Built-in risk categorization system
- **Employment Length:** Job stability impacts repayment capability

Technical Implementation

The solution utilizes a modern Python stack:

- **Core Libraries:** scikit-learn, XGBoost, pandas, numpy
- **Visualization:** matplotlib, seaborn, hvplot
- **Preprocessing:** MinMaxScaler, train_test_split
- **Evaluation:** ROC-AUC, confusion matrices, classification reports

Code is modular and production-ready with comprehensive evaluation metrics.

Conclusions & Future Work

This project successfully demonstrates the power of machine learning for credit risk assessment. The XGBoost model achieves excellent discrimination capability (ROC-AUC = 0.907), providing LendingClub with a robust tool for default prediction.

Future Enhancements:

- Advanced feature engineering using domain expertise
- Hyperparameter optimization through grid search
- Cost-sensitive learning to account for business impact
- Integration with real-time data pipelines
- Explainable AI techniques for regulatory compliance

The methodology and results establish a strong foundation for production deployment and continued model improvement.