EVALUATING LOAN DEFAULT RISK USING DECISION TREES AND RANDOM FORESTS

Executive Summary

This case study explains how classification algorithms—Decision Trees and Random Forests—can be used to predict the likelihood of loan default. Using a real-world-style financial dataset, students learn to preprocess data, select features, train models, and interpret predictions. The comparison of tree-based methods highlights the balance between model accuracy and explainability, giving students practical skills aligned with finance, banking, and credit analytics.

1. Introduction

Lending institutions must assess the risk of loan applicants defaulting. Machine learning models can assist by identifying patterns in past applicants' data. Decision Trees offer easy interpretability, while Random Forests boost accuracy through ensemble learning. This project teaches both, making it ideal for data science coursework, applied machine learning classes, or credit scoring challenges.

2. Problem Statement

Build and evaluate models that predict whether a loan applicant will **default (1)** or **repay (0)** based on their personal, financial, and loan attributes.

3. Dataset Overview

- **Source**: Simulated credit dataset (based on UCI/Loan Prediction data)
- **Observations**: 5,000 applicants
- **Features**:
    - age, income, employment_length, loan_amount, interest_rate,
    - credit_score, loan_term, home_ownership, previous_defaults,
    - loan_default (target: 0 = repaid, 1 = defaulted)

4. Methodology

Step 1: Preprocessing

- Imputed missing values in income and credit score
- Converted categorical variables (home_ownership, loan_term) using one-hot encoding
- Split data into train (70%) and test (30%)

Step 2: Model Building

a) Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
model_dt = DecisionTreeClassifier(max_depth=5)
model_dt.fit(X_train, y_train)
```

b) Random Forest

```
from sklearn.ensemble import RandomForestClassifier
model_rf = RandomForestClassifier(n_estimators=100)
model_rf.fit(X_train, y_train)
```

5. Results

Accuracy Comparison

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	84.2%	0.81	0.78	0.79
Random Forest	88.9%	0.87	0.84	0.85

Feature Importance (Random Forest Top 5)

1. Credit Score
2. Loan Amount
3. Income
4. Previous Defaults
5. Interest Rate

6. Visualization

- **Decision Tree Plot** (Gini index, max depth = 5)

- **Bar Chart** of feature importances
- **Confusion Matrix Heatmap**
- **ROC Curves** for both models (AUC: DT = 0.85, RF = 0.91)

7. Interpretation

- **Credit score** is the strongest predictor of default
- **Previous defaults** and **loan amount** increase risk
- **Random Forest** performs better but is less interpretable
- **Decision Tree** offers rule-based insights that are easier to explain to non-technical stakeholders

8. Conclusion

This project shows how classification algorithms can be used to assess financial risk. Students learn to build, compare, and interpret tree-based models while developing insights relevant to banking, fintech, and risk management.

9. Learning Outcomes for Students

- Train and evaluate decision tree models
- Understand ensemble techniques (Random Forest)
- Extract and explain feature importance
- Translate predictions into risk profiles and business decisions

10. Suggested Enhancements

- Apply cross-validation and hyperparameter tuning
- Compare with logistic regression or XGBoost
- Create a Flask web interface for input and scoring
- Implement SHAP for explainable AI in financial models

11. References

- Hastie, Tibshirani, & Friedman. *The Elements of Statistical Learning*
- Scikit-learn Docs: Decision Trees and Random Forests
- UCI ML Repository: Credit Risk Data
- Interpretable ML Book by Christoph Molnar