Customer Risk Weighting Prediction Using Machine Learning

1. Introduction

This project focuses on predicting customer Risk Weighting using machine learning techniques. Risk Weighting is a critical metric used by financial institutions to assess the likelihood of default and to make informed lending decisions. The objective is to build an end-to-end machine learning pipeline that cleans data, engineers features, trains multiple models, evaluates them, and generates reliable predictions.

2. Dataset Overview

The dataset consists of 3000 customer records containing demographic, financial, and behavioral attributes. The target variable is Risk Weighting, categorized into five classes ranging from low to high risk.

3. Data Understanding and Exploration

Initial exploration was conducted using methods such as info(), describe(), and head() to understand data types, distributions, and completeness. This step confirmed the absence of missing values and highlighted the mix of numerical and categorical features.

4. Data Cleaning

Irrelevant identifiers and redundant columns were removed to reduce noise. Data types were validated, and consistency checks ensured the dataset was suitable for modeling.

5. Feature Engineering

Three new features were engineered:

- Total Balance: Aggregation of checking, saving, and deposit accounts to reflect financial strength.

- Credit Utilization: Ratio indicating credit usage intensity.

- Loan to Income Ratio: Measure of debt burden relative to income.

These engineered features enhanced the model's ability to capture real-world financial behavior.

6. Feature Selection

Both original and engineered features were retained after experimentation showed improved performance. No severe multicollinearity issues were observed.

7. Target Variable Selection

Risk Weighting was selected as the target variable due to its business relevance and balanced class distribution compared to other potential targets.

8. Data Splitting

The dataset was split into training and testing sets using stratified sampling to preserve class distribution and avoid bias.

9. Encoding and Scaling

Categorical variables were encoded using OneHotEncoder. Numerical features were scaled using StandardScaler where required. All preprocessing steps were implemented within pipelines to prevent data leakage.

10. Model Building

Multiple models were trained:

- Logistic Regression

- Logistic Regression with SMOTE

- Random Forest Classifier

- Gradient Boosting Classifier

This comparative approach ensured robust model selection.

11. Model Evaluation

Models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Accuracy alone was not sufficient due to class imbalance, making ROC-AUC and F1-score critical metrics.

12. Best Model Selection

The Gradient Boosting model demonstrated the best overall performance with the highest ROC-AUC score and balanced predictions across classes.

13. Predictions

The final model was used to predict Risk Weighting for unseen customers. Probability outputs were analyzed to explain prediction confidence and support decision-making.

14. Business Interpretation

Predicted risk levels can be used by financial institutions to adjust lending policies, apply risk-based pricing, and flag high-risk customers for further review.

15. Conclusion

This project demonstrates a complete machine learning workflow, from data preprocessing to business interpretation. The resulting model provides actionable insights and a strong foundation for deployment.

16. Future Enhancements

Future work may include explainable AI techniques, cost-sensitive learning, dashboard visualization, and model deployment.