University of Technology and Engineering
Computer Science

Classification Project: Predicting Default Risk in Commercial Loans

Prof. Cristian López Del Alamo

September 5, 2025

Introduction to the Problem

The objective of this project is to build a **multiclass classification** model to predict the default risk level of clients applying for a commercial loan. The goal is not only to avoid high-risk clients, but to optimize the profitability of the loan portfolio, safely approving low and medium-risk clients while minimizing default losses. The database consists of **20,000 instances** and **35 features** per instance, classified as **0: Low**, **1: Medium**, and **2: High**.

Database

Download the database:

- Training Dataset
- Testing Dataset

Feature Names

The features are divided into three main categories:

1. **Financial Information (15 features)**
   - *annual_income*: Total client income in the last year.
   - *total_debt*: Amount of accumulated debt.
   - *income_to_debt_ratio*: Ratio between income and debt.
   - *loan_amount*: Amount of the requested loan.
   - *interest_rate*: Interest rate offered for the loan.
   - *open_credit_lines*: Number of active credit accounts.
   - *average_bank_balance*: Average balance in the client's accounts.
   - *per_capita_family_income*: Average income per family member.
   - *credit_bureau_score*: Official credit score.
   - *job_tenure_months*: Time the client has been in their current job.
   - *net_worth*: Value of assets minus liabilities.
   - *fixed_monthly_expenses*: Recurring monthly expenses.
   - *working_capital*: Difference between current assets and liabilities.
   - *monthly_savings_capacity*: Amount of money the client can save. *investment_income*: Income generated from investments.
2. **Payment History (10 features)**
   - *late_payments_last_6_months*: Number of delayed payments.
   - *historical_arrears_days*: Average days of payment delay.

- *credit_utilization_percentage*: Percentage of credit limit used.
- *on_time_payments_last_12_months*: Number of on-time payments.
- *historical_debts_paid_off*: Number of debts the client has fully settled.
- *maximum_payment_delay_days*: The longest payment delay the client has had.
- *number_of_closed_accounts*: Credit accounts that have been closed.
- *on_time_payment_ratio*: Ratio of on-time payments.
- *recent_credit_inquiries*: Number of times the credit history has been checked.
- *changes_in_payment_habits*: Variability in the client's payment patterns.

3. **Demographic and Behavioral Data (10 features)**
   - *age*: Client's age.
   - *educational_level*: Client's formal education level.
   - *marital_status*: Marital status.
   - *number_of_dependents*: Number of people financially dependent on the client.
   - *home_ownership*: Whether the client owns, rents, or has a mortgage.
   - *home_type*: Type of property.
   - *residence_tenure_months*: Time the client has lived at their current address.
   - *employment_sector*: Industry sector in which they work.
   - *number_of_jobs_last_5_years*: Employment stability.
   - *monthly_transaction_frequency*: Number of monthly bank transactions.

Project Challenges

This project is particularly challenging for students for the following reasons:

1. **Multiclass Classification**: The task is not binary, which requires a deeper analysis of the confusion matrix and per-class evaluation metrics.

2. **High Dimensionality**: With 35 features, students must consider *feature selection or reduction* to avoid overfitting.

3. **Cost of Errors**: A model that incorrectly classifies a "High" risk client as "Low" has a much higher financial cost than an error in another direction.

4. **Interpretability**: Students must go beyond accuracy and explain why their model made certain decisions, identifying the most important features.

Project Development Phases

1. **Data Analysis and Preprocessing**: Load the data, perform an exploratory analysis, and normalize or standardize the numerical features.

2. **Feature Selection and Reduction**: Use methods like PCA or feature importance from tree models to simplify the dataset.

3. **Modeling**: Train various classification algorithms, such as Logistic Regression, Support Vector Machines (SVM), and Random Forests.

4. **Evaluation and Optimization**: Use metrics like *Precision*, *Recall*, and *F1-Score* to evaluate performance. Perform cross-validation and hyperparameter optimization.

5. **Conclusions**: Present the findings, justify the model design decisions, and discuss the interpretability of the results.

Evaluation Rubric

The project evaluation will be based on a total of 20 points, distributed as follows:

Criterion	**Score**	**Performance Level**
1. Data Understanding and Preprocessing	**5 pts**	
Exploratory Data Analysis (EDA)	2 pts	**0-1.0 pts:** Limited or superficial EDA. **1.5 pts:** Adequate exploration and visualization. **2.0 pts:** Exhaustive EDA with analysis of relationships and clear visualizations.
Data Cleaning and Preparation	3 pts	**0-1.0 pts:** Inadequate preprocessing. **2.0 pts:** Adequate preprocessing. **3.0 pts:** Robust preprocessing and exploration of feature engineering.
2. Modeling and Training	**6 pts**	
Model Selection and Application	3 pts	**0-1.0 pts:** Use of a single model without justification. **2.0 pts:** Use of multiple models, but with flaws in the application. **3.0 pts:** Optimal implementation and justification of at least three models.
Dimensionality Handling	3 pts	**0-1.0 pts:** High dimensionality problem not addressed. **2.0 pts:** Feature selection/reduction techniques applied effectively. **3.0 pts:** Comparison and justification of the impact of different techniques.
3. Evaluation and Optimization	**5 pts**	
Use of Evaluation Metrics	2 pts	**0-1.0 pts:** Only accuracy is reported. **1.5 pts:** Use of appropriate metrics. **2.0 pts:** Deep analysis of the confusion matrix and the implications of errors.
Model Optimization	3 pts	**0-1.0 pts:** No hyperparameter optimization was performed. **2.0 pts:** Basic optimization with cross-validation. **3.0 pts:** Comparison of multiple optimization techniques that improved performance.
4. Report and Conclusions	**4 pts**	
Report Clarity and Structure	2 pts	**0-1.0 pts:** Disorganized or incomplete report. **1.5 pts:** Well-structured report, but without justification. **2.0 pts:** Clear, logical report that follows the requested academic format.
Business Interpretation	2 pts	**0-1.0 pts:** Superficial interpretation of the results. **1.5 pts:** The results are interpreted in the context of the problem. **2.0 pts:** A detailed analysis of the model's interpretability is performed, suggesting business strategies.

Report Submission Format

For the final report submission, groups must use the LaTeX template from the **Institute of Electrical and Electronics Engineers (IEEE)**. This format ensures a professional and standardized presentation, crucial for communication in the scientific and technical fields.

- **Link for downloading the IEEE LaTeX template**: `https://www.ieee.org/conferences/publishing/templates.html`