Credit Risk Prediction Using the German Credit Dataset

CS 4120 | Machine Learning, Data Mining

Project Proposal

Ditthi Chatterjee & Mohammed Sahm

1. Problem and Motivation

Credit risk assessment is a critical task for financial institutions. Accurately predicting whether a client will default or repay a loan allows banks to minimize losses and optimize lending decisions. Individuals benefit indirectly, as proper risk assessment leads to fairer loan offers and reduced interest rates for low-risk borrowers.

Beyond institutional use, making such models more accessible to the general public can empower individuals to better understand how financial behavior influences their creditworthiness. With clear feedback, people can learn which spending or repayment habits strengthen their credit profile, ultimately improving their credit scores. Higher credit scores not only increase access to loans and credit products but also lower borrowing costs, creating long-term financial stability. In this way, accessible risk models can serve as a practical tool for improving financial literacy, encouraging healthier spending habits, and promoting more equitable participation in the financial system.

This project is motivated by the need for reliable and interpretable machine learning models for credit risk classification and regression-based risk scoring.

---

2. Dataset Description

The dataset used in this project is the German Credit dataset provided by Prof. Dr. Hans Hofmann, University of Hamburg. It contains information about 1,000 individuals applying for credit, with 20 attributes (7 numerical and 13 categorical) describing financial and personal details.

- Source: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
- License: Creative Commons Attribution 4.0 International (CC BY 4.0)
- Number of instances: 1000
- Number of attributes: 20
- Data types: Numerical (Integer), Categorical (Binary)
- Missingness: 0% throughout the dataset
- Sensitive attributes: Sex, Foreign worker (No data is tied to a name/other identifiable information – fully anonymous)

Table 1 – Dataset Snapshot

Attribute Name	Data Type	Description	Units	Notes / Categories
Status of existing checking account	Categorical	Checking account status	-	A11: < 0 DM; A12: 0 ≤ … < 200 DM; A13: ≥ 200 DM / salary assignment ≥ 1 year; A14: no checking account
Duration	Numerical	Duration of credit	months	-
Credit history	Categorical	Past credit repayment record	-	A30: no credits taken / all paid; A31: all credits at this bank paid back duly; A32: existing credits paid duly till now; A33: delay in past; A34: critical/other credits
Purpose	Categorical	Purpose of credit	-	A40: car (new); A41: car (used); A42: furniture/equipment; A43: radio/TV; A44: domestic appliances; A45: repairs; A46: education; A48: retraining; A49: business; A410: others
Credit amount	Numerical	Amount of credit requested	Deutsche Marks (DM)	-

Savings account/bonds	Categorical	Status of savings/bonds	-	A61: < 100 DM; A62: 100 ≤ … < 500 DM; A63: 500 ≤ … < 1000 DM; A64: ≥ 1000 DM; A65: unknown/no savings
Present employment since	Categorical	Employment duration	years	A71: unemployed; A72: < 1 year; A73: 1–4 years; A74: 4–7 years; A75: ≥ 7 years
Installment rate	Numerical	Installment rate as % of disposable income	%	-
Personal status and sex	Categorical	Marital status and sex	-	A91: male, divorced/separated; A92: female, divorced/separated/married; A93: male, single; A94: male, married/widowed; A95: female, single
Other debtors / guarantors	Categorical	Presence of co-debtors or guarantors	-	A101: none; A102: co-applicant; A103: guarantor
Present residence since	Numerical	Years living at current residence	years	-
Property	Categorical	Type of property owned	-	A121: real estate; A122: savings agreement/life insurance; A123: car or other; A124: none/unknown
Age	Numerical	Age of applicant	years	-
Other installment plans	Categorical	Other existing installment plans	-	A141: bank; A142: stores; A143: none
Housing	Categorical	Housing situation	-	A151: rent; A152: own; A153: for free
Number of existing credits	Numerical	Number of credits at this bank	count	-
Job	Categorical	Employment type	-	A171: unemployed/unskilled non-resident; A172: unskilled resident; A173: skilled employee/official; A174: management/self-employed/highly qualified
Number of dependents	Numerical	People financially supported	count	-
Telephone	Categorical (Binary)	Presence of telephone	-	A191: none; A192: yes, registered under customer's name
Foreign worker	Categorical (Binary)	Foreign worker status	-	A201: yes; A202: no
Credit Risk (Target)	Categorical (Binary)	Creditworthiness label	-	**Target: Credit Risk (`1 = Good, 2 = Bad`)** **Class Distribution: 700 good (70%)** **300 bad (30%)**

---

3. Tasks

Classification Task: Predict Credit Risk (Good / Bad). Derived directly from the dataset's target label. Feasible due to clear labeling and sufficient instances per class.

Regression Task: Predict Credit Amount (units: DM). Feasible because the target is numerical and influenced by other financial and demographic attributes.

---

4. Metrics Plan

- <u>Classification:</u> Accuracy, F1-score, ROC-AUC
- <u>Regression:</u> Mean Absolute Error (MAE), Root Mean Squared Error (RMSE)

Metrics will be calculated on a held-out test set with appropriate cross-validation to ensure robustness.

---

5. Baseline Plan (Classical ML)

We will establish classical ML baselines for both tasks:

Task	Model 1	Model 2
Classification	Logistic Regression	Decision Tree
Regression	Linear Regression	Decision Tree Regressor

---

6. Reproducibility Plan

- Dependency pinning: Using requirements.txt with fixed versions.
- Random seed setting: Fix random seeds (NumPy, scikit-learn) so experiments produce consistent results across runs.
- MLflow tracking: Log parameters, metrics, and artifacts for all experiments.
- Git repository: Initialized with structured folders: /data /notebooks /src /models /reports README.md requirements.txt

Table 2 – Planned Models and Metrics

Task	Model	Hyperparameters / Notes	Metrics to Report
Classification	Logistic Regression	Default; may tune C, penalty=L2	Accuracy, F1-score, ROC-AUC
Classification	Decision Tree Classifier	Tune `max_depth`, `min_samples_split`, `min_samples_leaf`	Accuracy, F1-score, ROC-AUC
Regression	Linear Regression	Default	MAE, RMSE
Regression	Decision Tree Regressor	Tune `max_depth`, `min_samples_split`, `min_samples_leaf`	MAE, RMSE

Notes:

- We did not include Naïve Bayes because it is not directly suited to continuous-valued features without additional preprocessing or discretization, which is outside the scope of our baseline plan. Logistic Regression and Decision Trees already provide strong, interpretable baselines that align with both the dataset characteristics and course requirements.
- While not required, we may use Dummy Classifiers/Regressors (e.g., predicting the majority class or mean) as trivial baselines to benchmark whether our chosen models provide meaningful improvements.

Submission Package:

- PDF named proposal_GermanCredit-G1.pdf
- Git repository link with initialized structure and code: https://github.com/ninichatterjee/german-credit