Group B8 project report

Kaur Vadi

Tregert Värv

Business understanding

Background

Credit risk assessment is a critical process for financial institutions. Incorrect risk evaluation leads to financial losses, while overly strict approval criteria decrease potential profit and customer base. Our project uses a Kaggle dataset (from Playground Series S5E11) based on the synthetic "Loan Prediction Dataset". The dataset contains borrower demographics, financial information, loan information and credit history. With this information, we aim to build a model that predicts whether a borrower will repay a loan or default.

Business goals

Although this is not a real business case, we perform the business understanding as if the end-user is a financial institution (e.g. a bank) issuing personal loans.

1. Develop an automated loan-risk evaluation model.
2. Reduce potential financial losses by identifying risky borrowers early.
3. Improve loan approval decision-making efficiency.

Business success criteria

A successful outcome will be achieved if:

- The model predicts loan repayment status with high accuracy (or AUC score).
- False-positive approval of risky customers is significantly reduced.

We aim for **minimum AUC > 0.90**

Inventory of resources

Our project team consists of two members with basic experience in Python, machine learning and exploratory data analysis gained from course practical sessions. We use a synthetically generated dataset provided by Kaggle (Playground Series S5E11), which contains all relevant features required for credit-risk prediction: demographics, financial attributes, loan details and credit history. The tools available for this project include Python, Jupyter Notebook, pandas, NumPy, scikit-learn and Kaggle's submission environment.

Requirements, assumptions and constraints

We assume that the dataset is internally consistent and suitable for building predictive models, as it is fully cleaned and includes no missing values. A key constraint is that the dataset is synthetic and therefore may not fully reflect real-world lending behavior. Another practical constraint is that our analysis must remain within the scope of the course timeline and tools taught in the practical sessions.

Risks and contingencies

Since the dataset is synthetic it may be biased towards modeling unrealistic behaviour. Therefore we need to evaluate the fairness of the results of our models as well. Another risk is timeline - we may not be able to construct the models fast enough to enter the Kaggle competition since the competition deadline does not align with the project deadline. However, in the general picture, this is no obstacle since we plan to evaluate our models ourselves as well, in addition to the official Kaggle competition submission.

Terminology

AUC: Area Under ROC Curve — evaluation metric for binary classification.

Feature: attribute of the borrower/loan used for prediction.

Default: borrower fails to repay the loan fully

Costs and benefits

This project does not involve financial cost, as all tools and datasets are freely available. The main investment is team time spent on analysis. The benefit is gaining practical experience with credit-risk modelling, synthetic datasets, and model evaluation in a Kaggle competition setting.

Data-mining goals

- Build a binary classification model predicting loan status (paid vs not paid).
- Compare multiple models: Logistic Regression, Random Forest, KNN, SVM, Gradient Boosting, XGBoost
- Analyse which borrower/loan features most influence repayment probability.

Data-mining success criteria

1. High generalization performance measured by AUC, accuracy, precision, recall.
2. Clear insights and visualizations explaining what drives loan repayment

Data understanding

Gathering, describing data

The data for our project is provided by Kaggle. This dataset is used for the Kaggle Playground Series Season 5 Episode 11 competition. The original dataset - “Loan Prediction Dataset” - was synthetically created by Nabiha Zahid (https://www.kaggle.com/datasets/nabihazahid/loan-prediction-dataset-2025). This dataset was then modified by Kaggle and their deep learning model to suit the needs of their competition. The dataset itself is completely cleaned and organized and includes no missing values. The data is also fully numeric and in a standardized format. Therefore, we don’t have to worry about cleaning the data and can focus solely on building machine learning models. The dataset comprises multiple sections: borrower’s demographics, financial information, loan information and borrower’s credit history. This is exactly the kind of information necessary to predict whether a borrower will pay back their loan or default. Link to Kaggle competition: https://www.kaggle.com/competitions/playground-series-s5e11

Exploring data

In order to make more sense of our data, we plan to create distribution plots and histograms: to understand differences in income, loan size and interest rate. We also plan to create a pairplot of key features vs default outcome. If our dataset turns out to be imbalanced, SMOTE or class-weighting may be needed. Potential insights we expect from exploring our data:

1. Higher interest and debt-ratio likely correlate with higher default rate.
2. Borrowers with a worse credit score are more likely to default

Verifying the data

Our dataset is synthetic and therefore may not represent real life patterns exactly. Since this dataset is used in a Kaggle competition we expect patterns to emerge from the data but they may be artificially amplified. Another limitation of this dataset is the absence of macroeconomic trends in the data.

The dataset is structured into:

- **train.csv** → includes features + target column *loan_default*
- **test.csv** → used for Kaggle submission
- **sample_submission.csv**

We expect to use most fields for predictive modeling except a redundant ID column.

Project plan

1. Data loading and initial structuring — 6h (3h + 3h)
    - We will load the Kaggle loan dataset into Python and confirm that all fields are correctly read. Since the dataset is already cleaned and contains no missing values, the main focus is to ensure that the data types are correct and that the unnecessary fields are removed.
2. Initial data overview— 10h (5h + 5h)
    - To understand the dataset, we will review summary statistics and basic feature distributions. This step helps us recognise key patterns in demographics, financial indicators and credit history, and gives us a general sense of which variables may influence loan repayment.
3. Feature analysis and preparation— 12h (6h + 6h)
    - We will examine all available features to understand their roles in predicting loan repayment. This includes interpreting numeric ranges, exploring categorical values, and preparing a consistent feature set that can be used across all machine-learning models.
4. Model building and validation — 18h (9h + 9h)
    - We train and compare several models that we have learned in the practice sessions. To measure the accuracy, we will evaluate all

models using ROC AUC, which is the required metric in the Kaggle competition.

5. Interpretation and final report — 14h (7h + 7h)
    - We will interpret model results, analyse which features influence repayment probability, create visualisations, and compile the final project report. Finally, we will generate predictions for the Kaggle test set and submit them as a CSV file.

Total time per teammate: 30h

Link to the GitHub repository of our project:
https://github.com/Kaurx14/IDS_2025_project