Machine Learning on Loan Approval Prediction

Qingyang Cheng
Brown University - Data Science Institute

DATA1030 Course Project

Date of Presentation: October 21st, 2024
Github repository: https://github.com/CQY114/data1030_fall2024_final_project.git

Project Overview

Loan Approval Prediction

- Binary classification
- “Will a new loan application be approved?”
- {income, loan amount, interest rate, etc.} —> **loan status**
- Importance: efficiency, insights

Link to Kaggle: https://www.kaggle.com/datasets/itshappy/ps4e9-original-data-loan-approval-prediction/data

Exploratory Data Analysis

Data Overview

- # data points: 32,581
- # features: 11
  - 7 numerical
  - 3 categorical (1 of them binary Y/N)
  - 1 ordinal
- Target: loan_status (binary)
  - 1: approved
  - 0: not approved

Exploratory Data Analysis

Unreasonable Data

- Both are in years
- AGE < LENGTH OF EMPLOYMENT (?!)

	person_age	person_emp_length
0	22	123.0
210	21	123.0

Exploratory Data Analysis

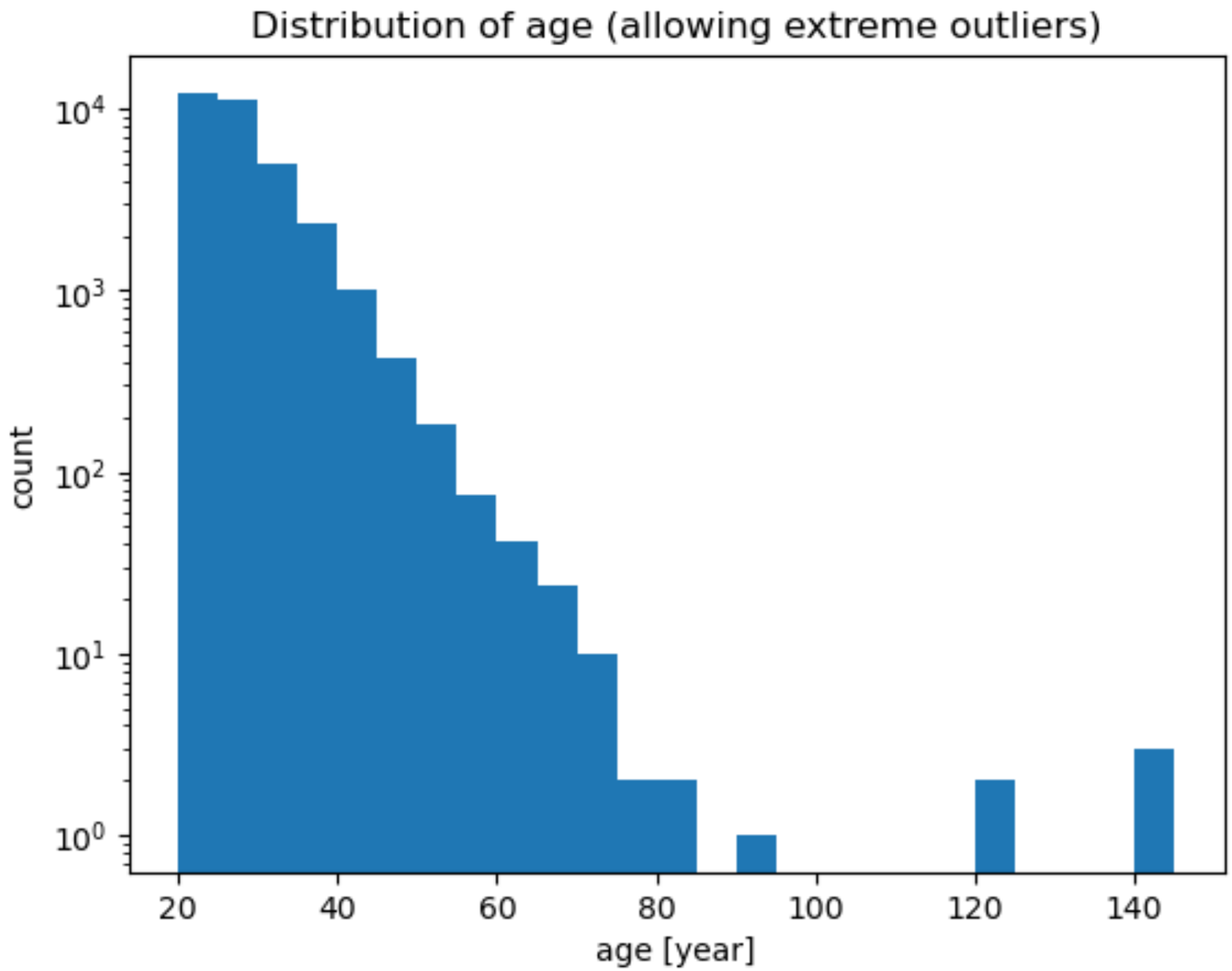
Unreasonable Data

- More...

person_age	person_emp_length
81	4.0
183	4.0
575	2.0
747	7.0
32297	12.0

Exploratory Data Analysis

Unreasonable Data



Exploratory Data Analysis

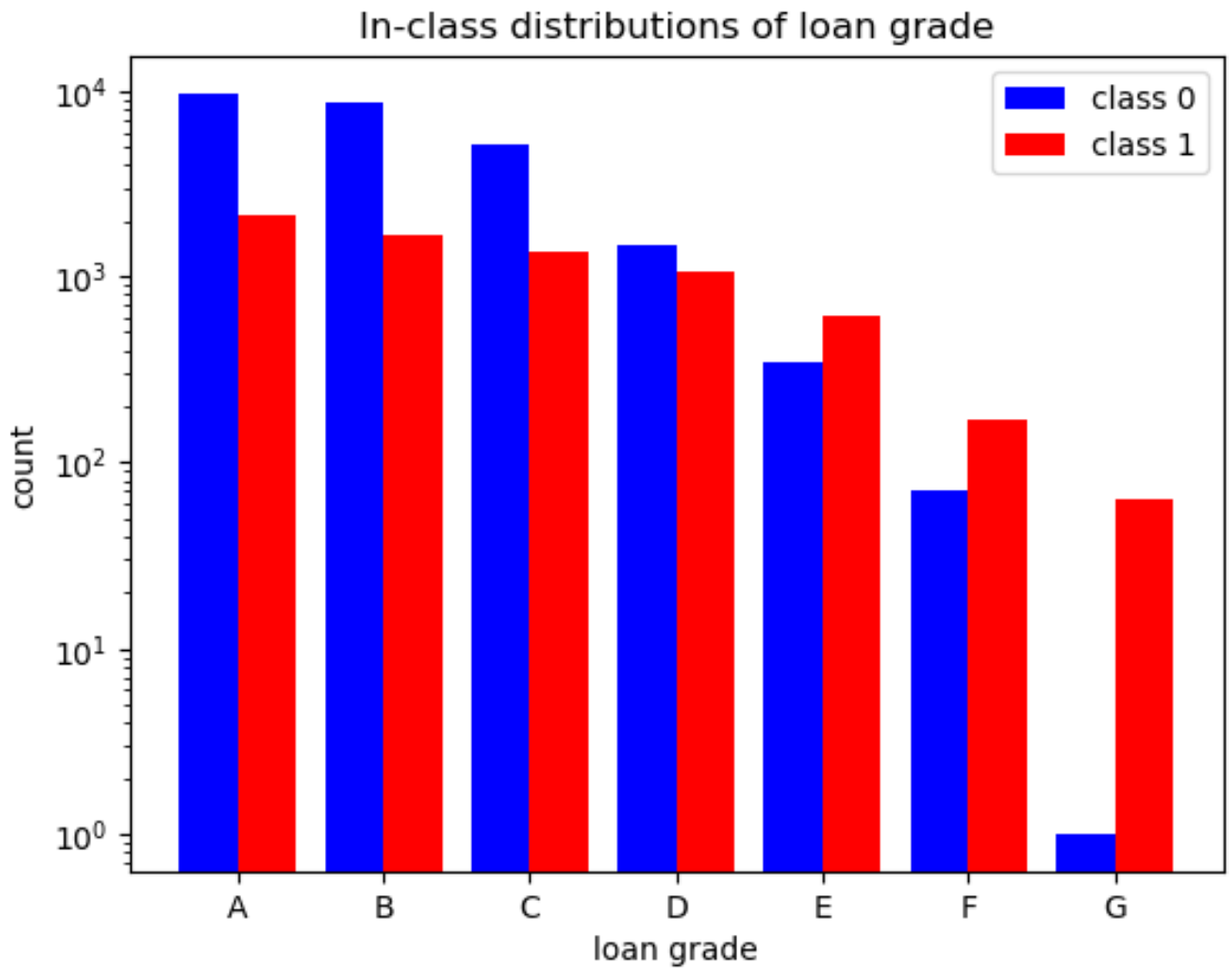
Distributions

- Delete 7 rows
- # data points: 32,574
- 7,107 points in class 1 (≈ 21.8%)

Exploratory Data Analysis

Distributions

- 7,107 points in class 1 (≈ 21.8%)



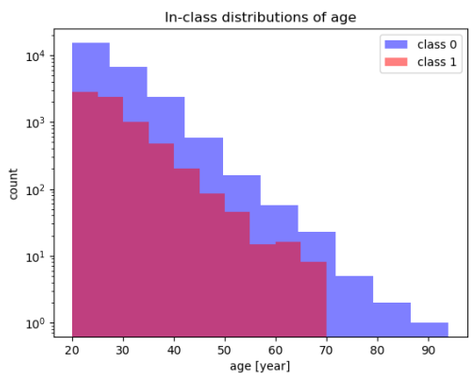
Exploratory Data Analysis

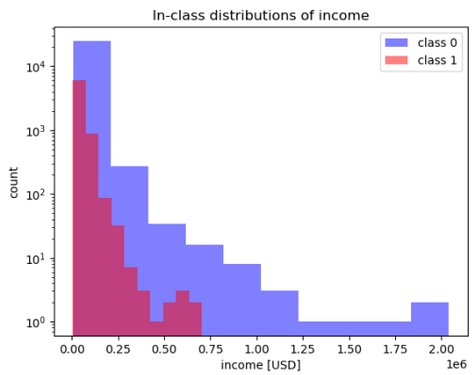
Distributions

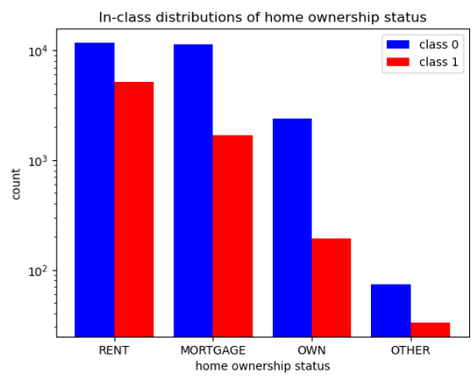
- 7,107 points in class 1 (≈ 21.8%)

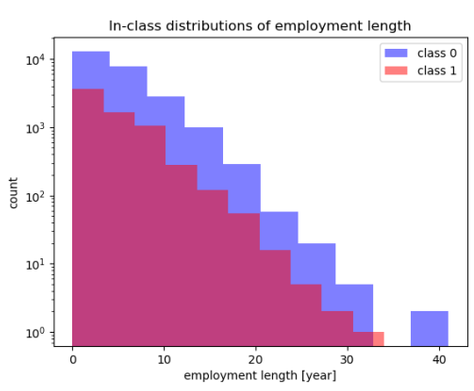
Exploratory Data Analysis

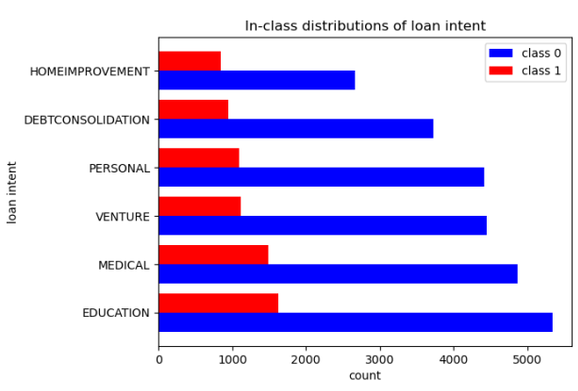
Distributions

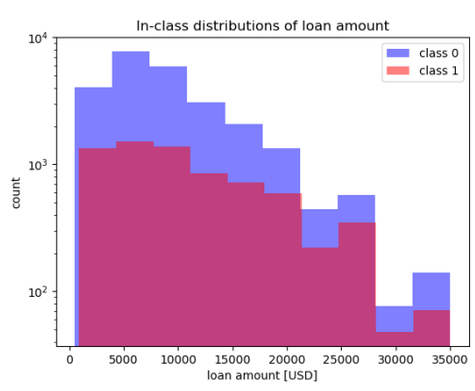


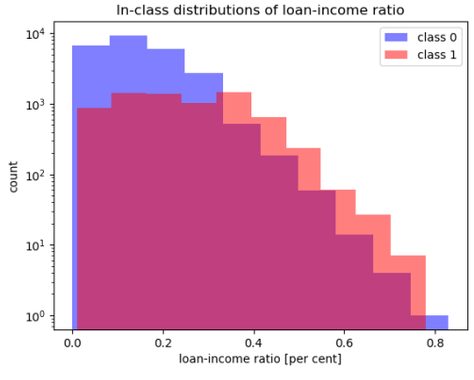


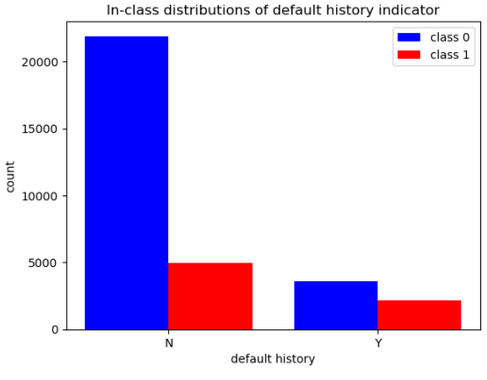












Exploratory Data Analysis

Missing Values

- # data points with missing values: 3,942 (≈ 12.1%)
- Features with missing values
  - employment length (numerical): 895
  - loan interest rate (numerical): 3,115

Preprocessing

Splitting

- 7,107 points in class 1 (≈ 21.8%) —> stratified splitting
- Ratio of splitting
  train : validation : test = 0.9 : 0.05 : 0.05
- Result of splitting
  # train = 29,316
  # validation = 1629
  # test = 1629
- ≈ 21.8% of points in class 1 for all 3 subsets

Preprocessing

Transformers

- Numerical features — standard scalers
- Categorical features — one-hot encoder
- *Binary feature — mapping {'Y': 1, 'N': 0}*
- Ordinal feature — ordinal encoder

Preprocessing

Transformers

- Resulting # features: 19
  - 7 numerical
  - 2 categorical
    - person_home_ownership (4-category)
    - loan_intent (6-category)
  - 1 binary
  - 1 ordinal

Q&A