Midterm project presentation

Loan Default Dataset Analysis

GitHub repository:
https://github.com/Bail111/Loan_Default-Analysis.git

REPORTER： Qinjunjie Pu

October 25th

01

troduction

Introduction

This project aims to predict whether a loan will default.

Minimizing financial losses and optimizing lending decisions.

Classification problem

Data Source: kaggle

Data Collection: past data on the loan borrowers

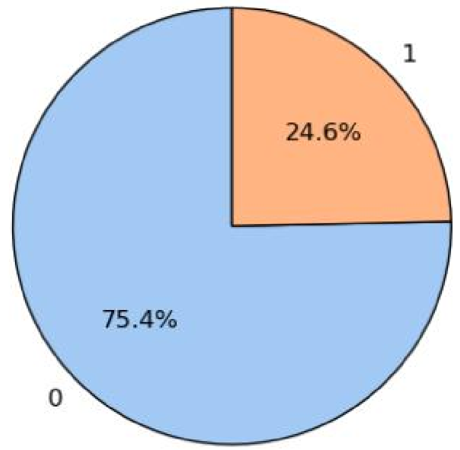
02

Exploratory Data Analysis

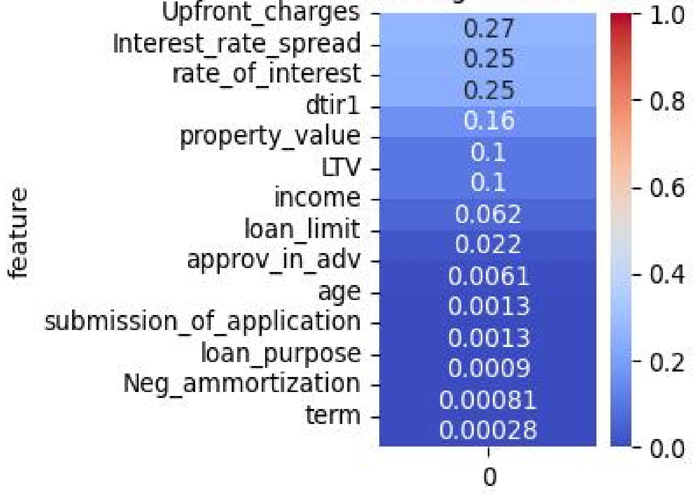
EDA

Brief description of dataset

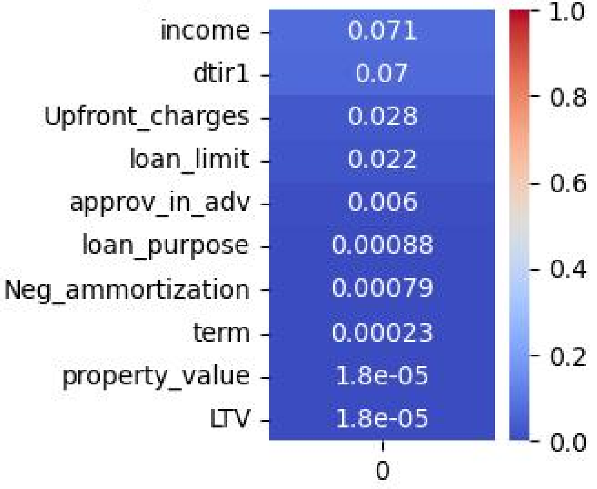
```
row: 148670
col: 34

12 continuous features
21 categorical features
```



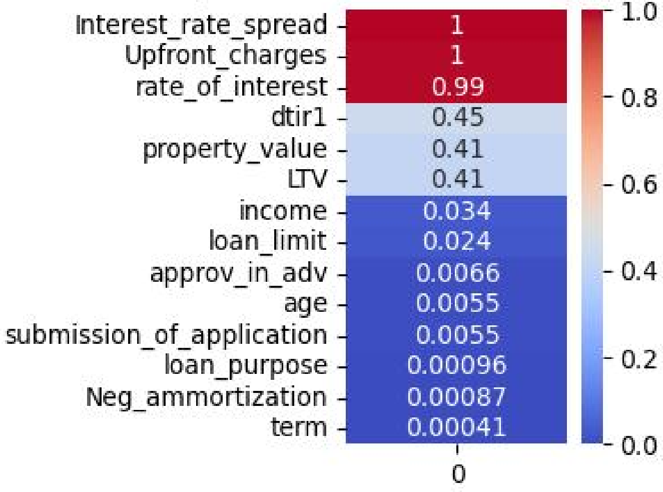
Loan Default Status Distribution



Overall Missing Value Distribution



Missing Value Distribution for Non-Default Data

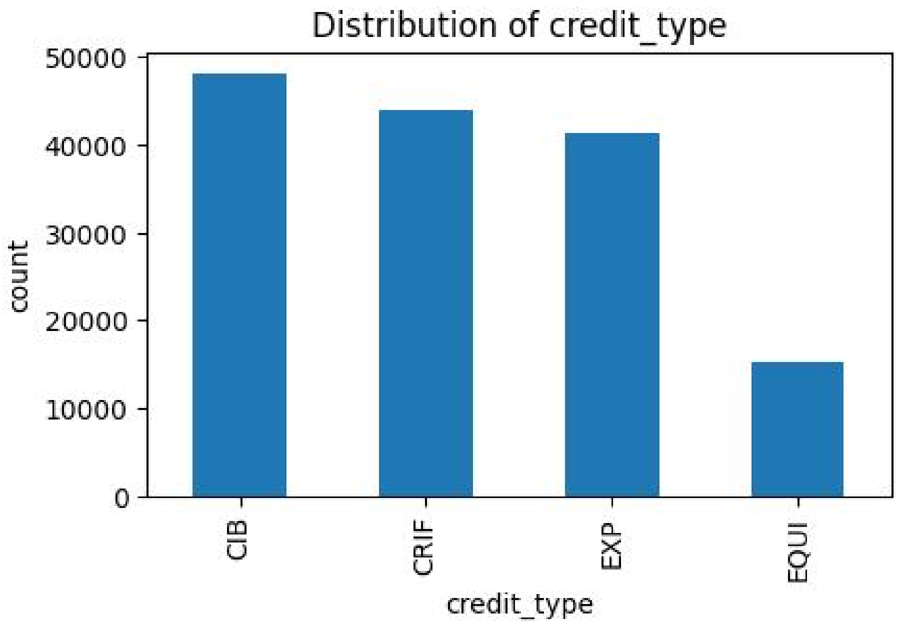


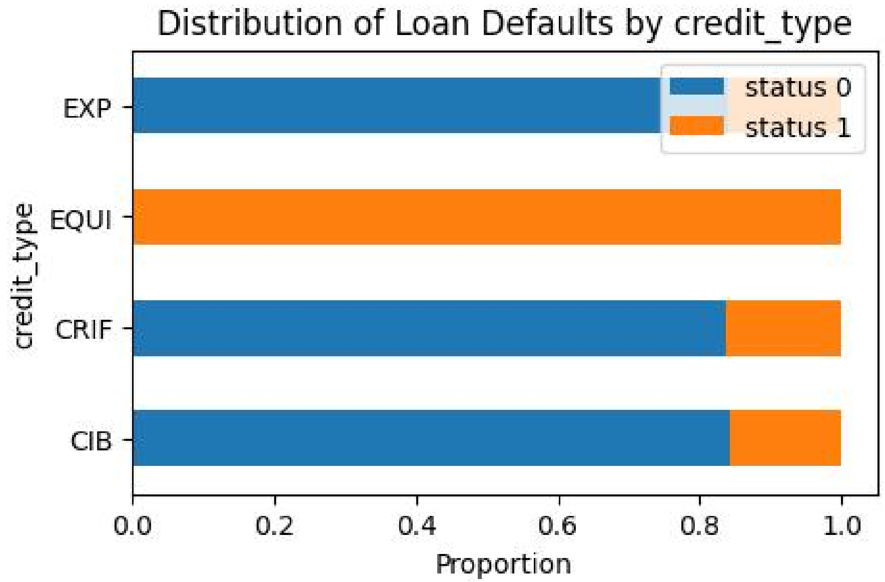
Missing Value Distribution for Loan Default Data

EDA

Analysis for categorical features

credit_type : applicant's type of credit

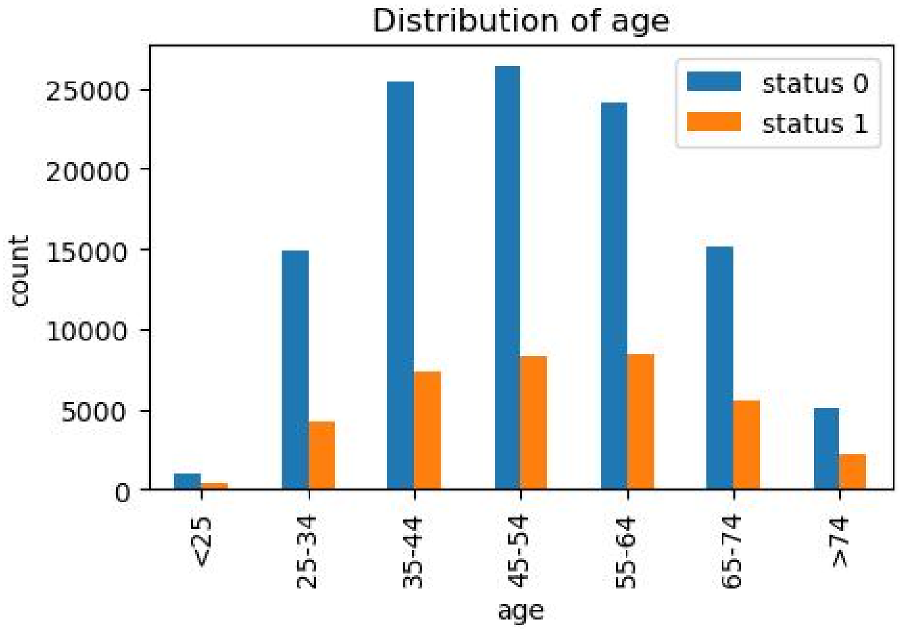


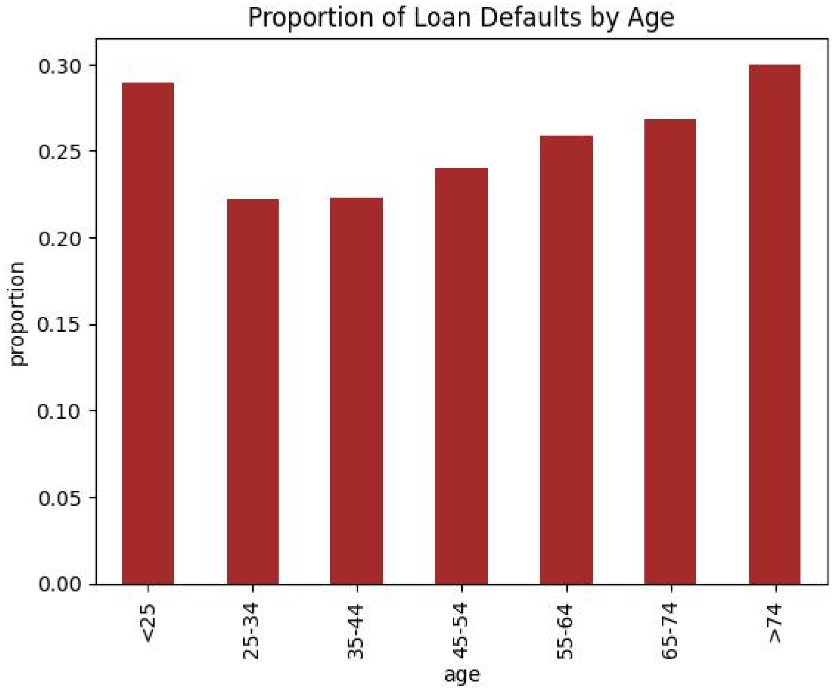


EDA

Analysis for categorical features

age: the age of the applicant

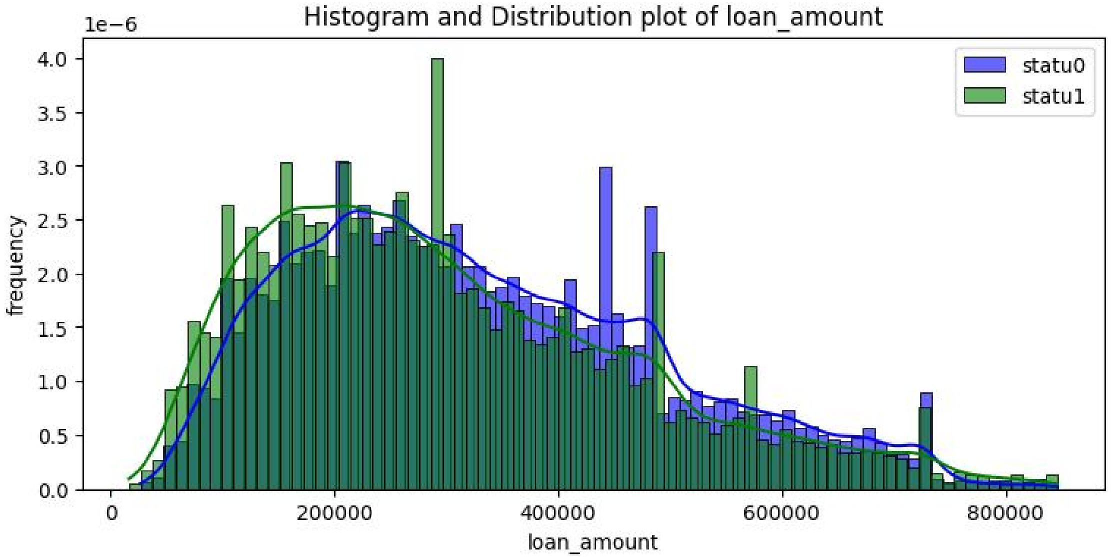

Distribution of age


Proportion of Loan Defaults by Age

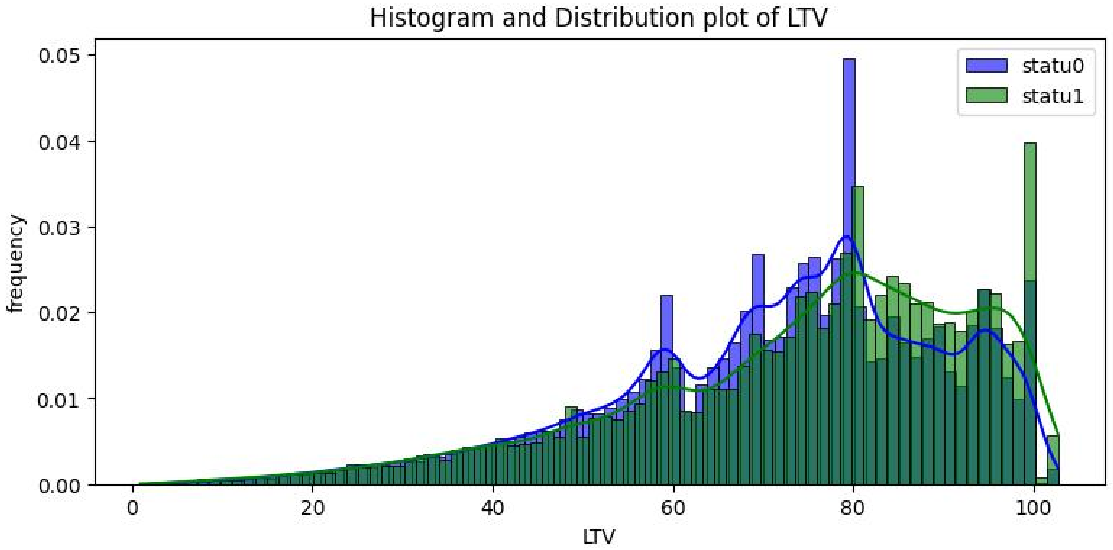
EDA

Analysis for continuous features

loan_amount: amount of money being borrowed

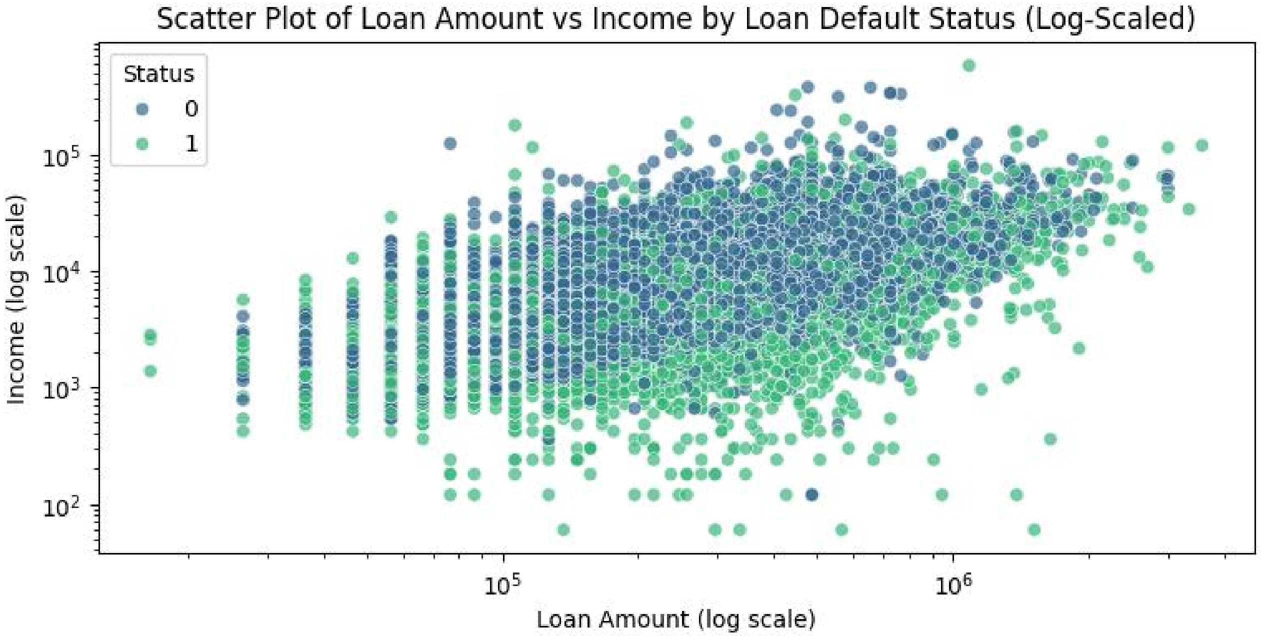


LTV: loan-to-value ratio, calculated as the loan amount divided by the property value



EDA

Analysis for continuous features



Income: applicant's annual income

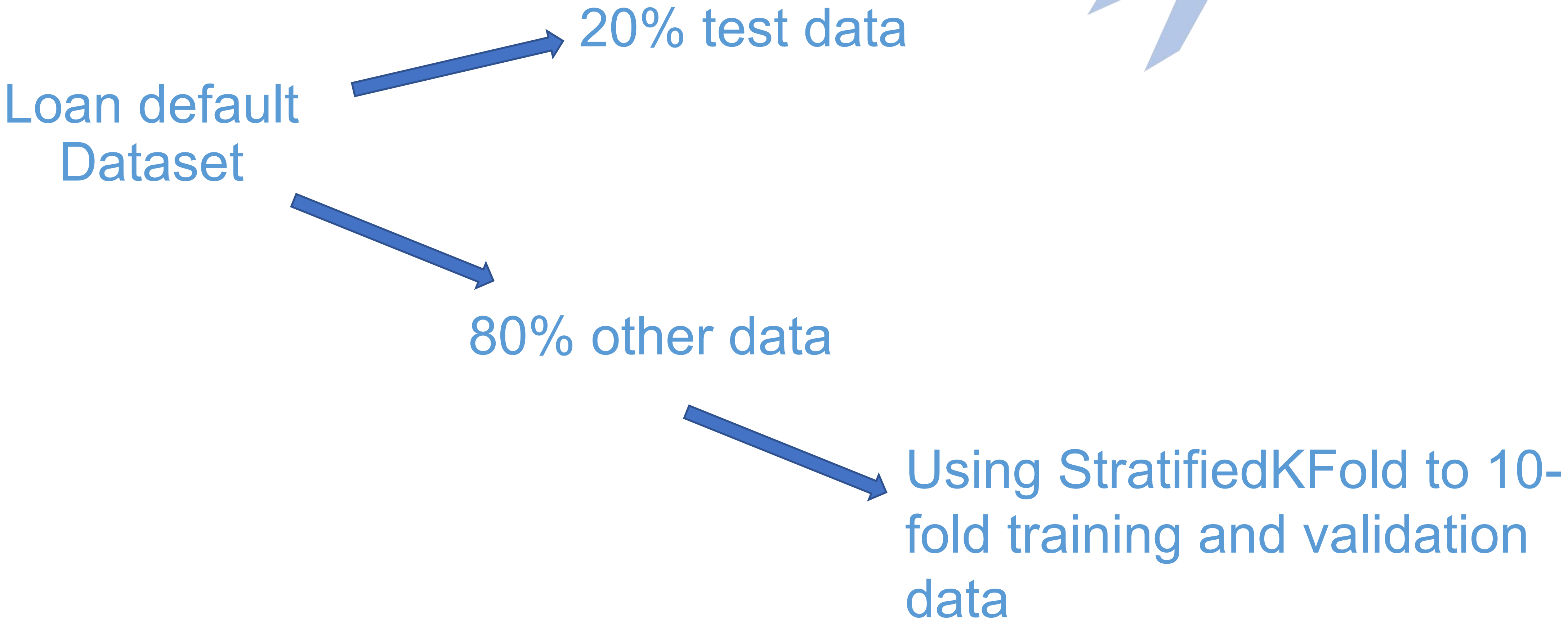
loan_amount: amount of money being borrowed

03

Splitting and preprocessing

Splitting and preprocessing

Splitting method



Loan default Dataset

20% test data

80% other data

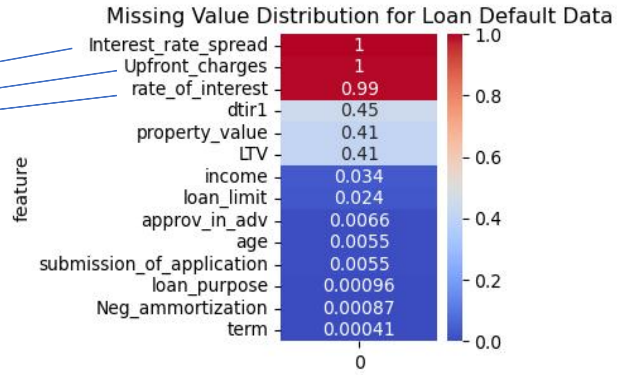
Using StratifiedKFold to 10-fold training and validation data

Splitting and preprocessing

Preprocessing

ID, year (2019 for all data)

1. Delete features



2. For categorical and ordinal features:

Impute missing value with ‘missing’

Binary encoding	e.g. business_or_commercial (‘nob/c’ to 0, ‘b/c’ to 1)
Onehot encoding	e.g. region
Ordinal encoding	e.g. age ('<25’, '25-34’, '35-44’, etc.)

3. For continuous features:

Minmax scalar	e.g. credit_score
Standard scalr	e.g. income

Before preprocessing: 33 features
After preprocessing: 53 features

end

Thanks for listening