Loan Default Prediction

for financial loan services

Jing Xu, Brown University DSI, 10/23/2024

GitHub:https://github.com/Jing-Xu1223/DATA1030-Project

Introduction

Imagine I am a Data Scientist working at a Financial loan service:

- One of my primary objectives for my company is to decrease payment defaults and ensure that all individuals are paying back their loans as expected.

- In order to do this efficiently and systematically, I would employ machine learning models to predict which individuals are at the highest risk of defaulting on their loans, based on their personal demographics and income summary.

- Thus, proper interventions can be effectively deployed to the right audience.

Choosing this project would enable me get more acquainted with the way data science operates within financial institutions.

Dataset Overview

This is a **Binary Classification** problem!

The Target Variable "Default" contains two classes:

Class 0: The borrower repays the amount—---no loan default
Class 1: The borrower failed to make payments—-- resulted in loan default.

01 Large Dataset: 255,347 rows and 18 columns (1 target + 16 features + 1 unique identifier)

02 IID Dataset: Each column represents a unique individual with his demographics and loan outcomes

03 No missing Value!

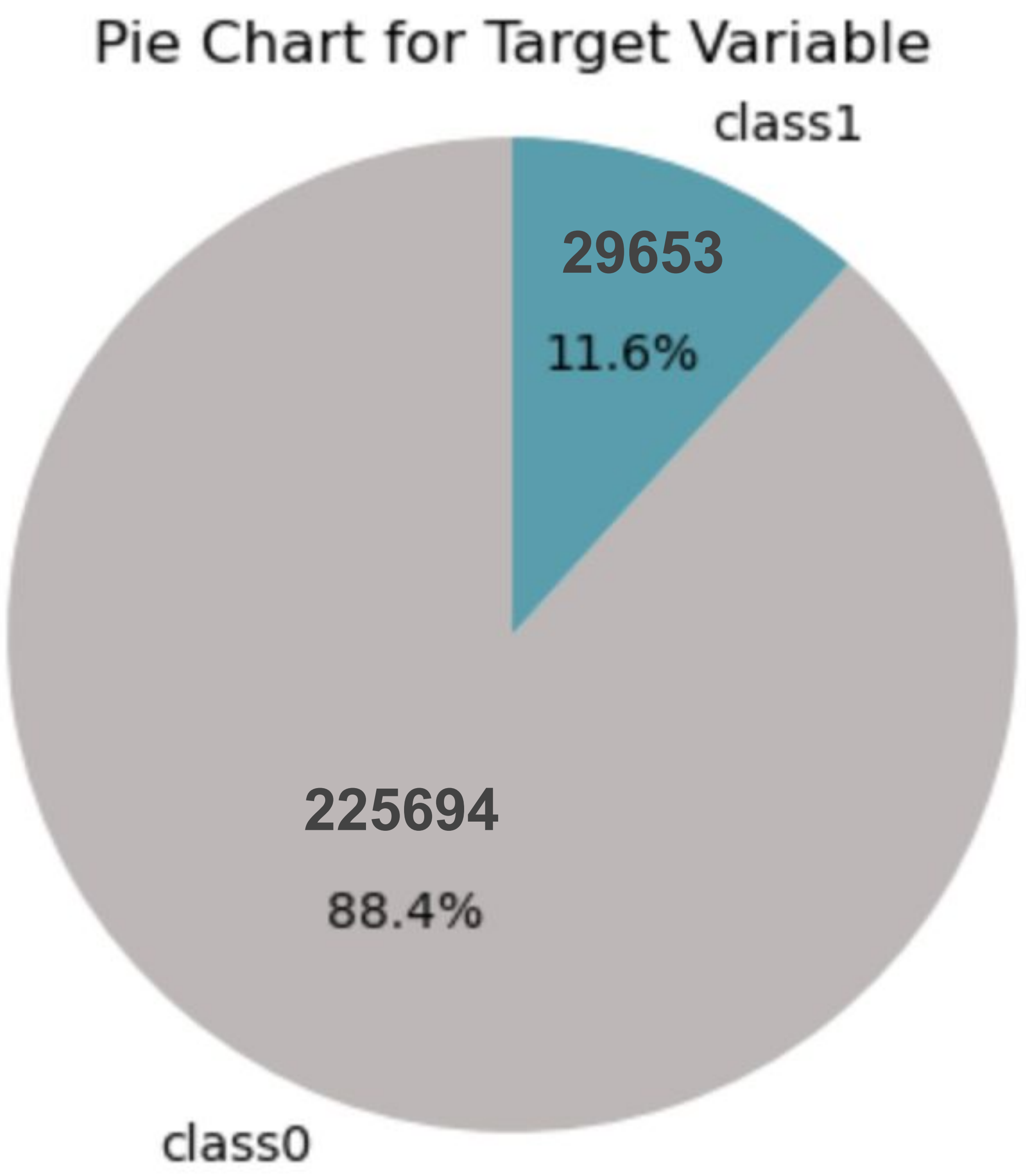
Dataset Collection

Dataset is available on Kaggle:Loan Default Prediction
This dataset is collected by Coursera Project Network:Loan Default Prediction Coding Challenge, which includes a sample of individuals who took financial loans in 2021.

EDA Part I: Target&Feature Column Analysis

01

Target Variable “Default”



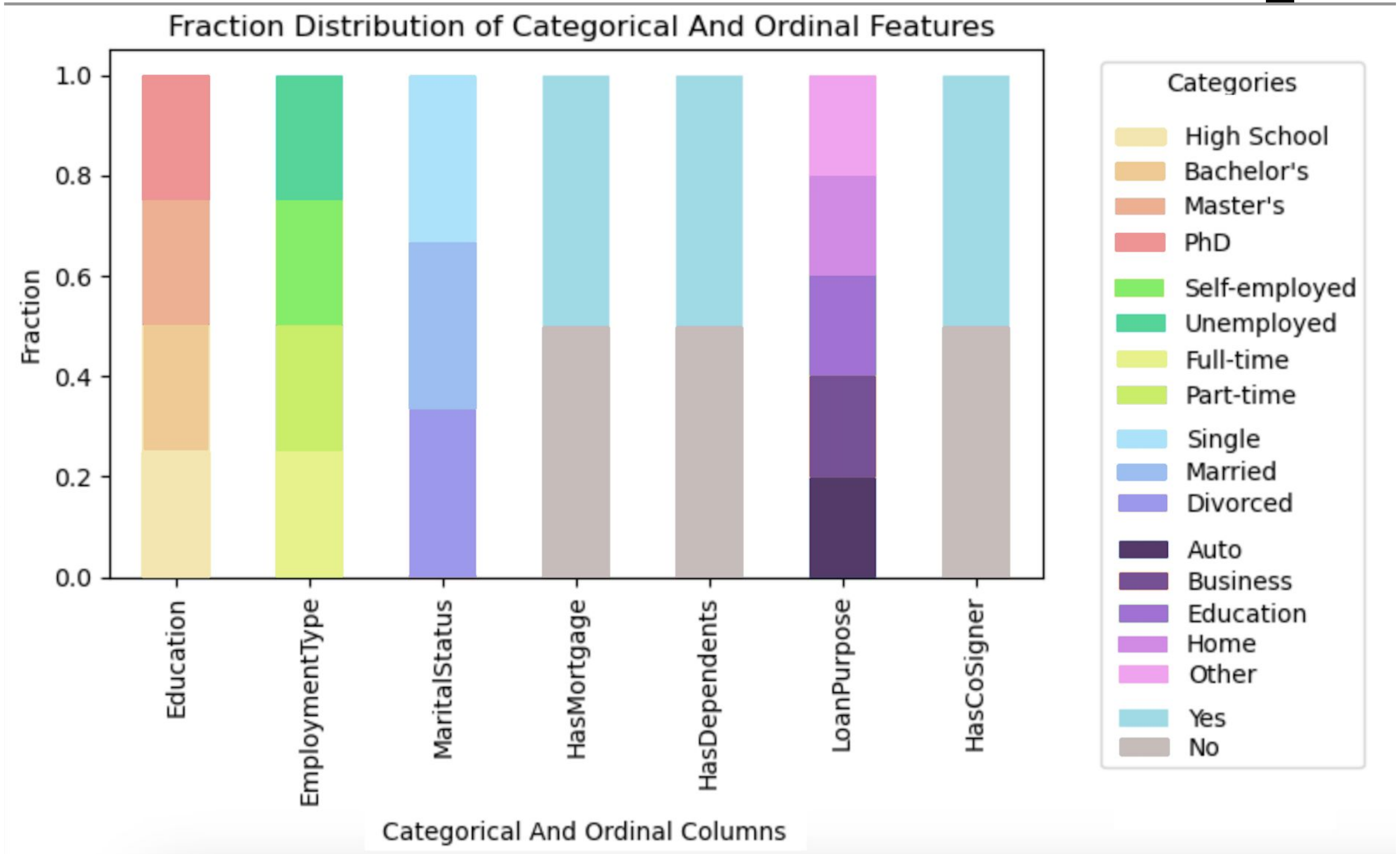
The majority of customers are likely to make loan payments.
Highly Imbalanced!(Stratify when splitting)

01

Target Variable “Default”

02

Categorical&Ordinal Features



All categorical and ordinal features are uniformly distributed.

01

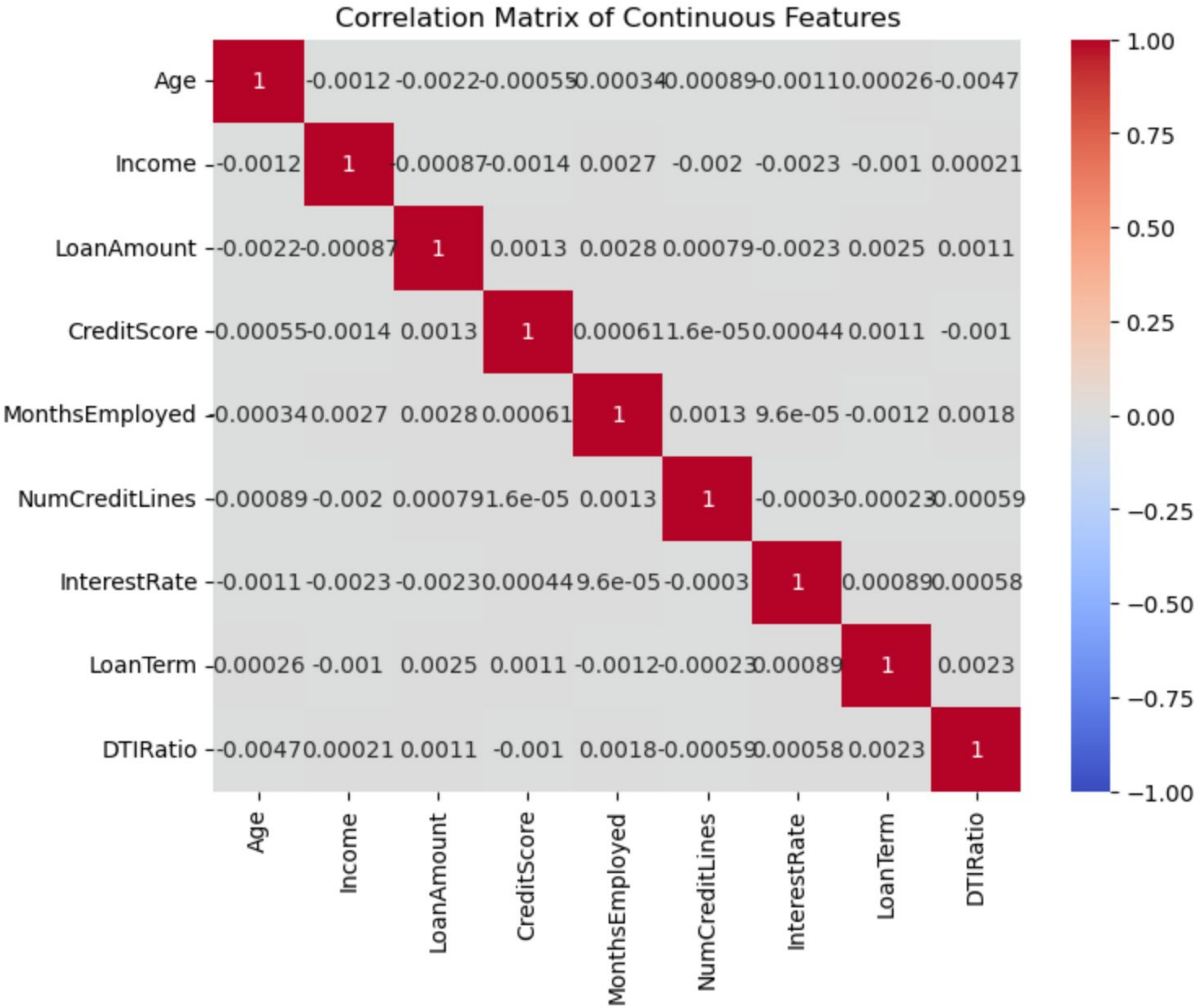
Target Variable “Default”

02

Categorical&Ordinal Features

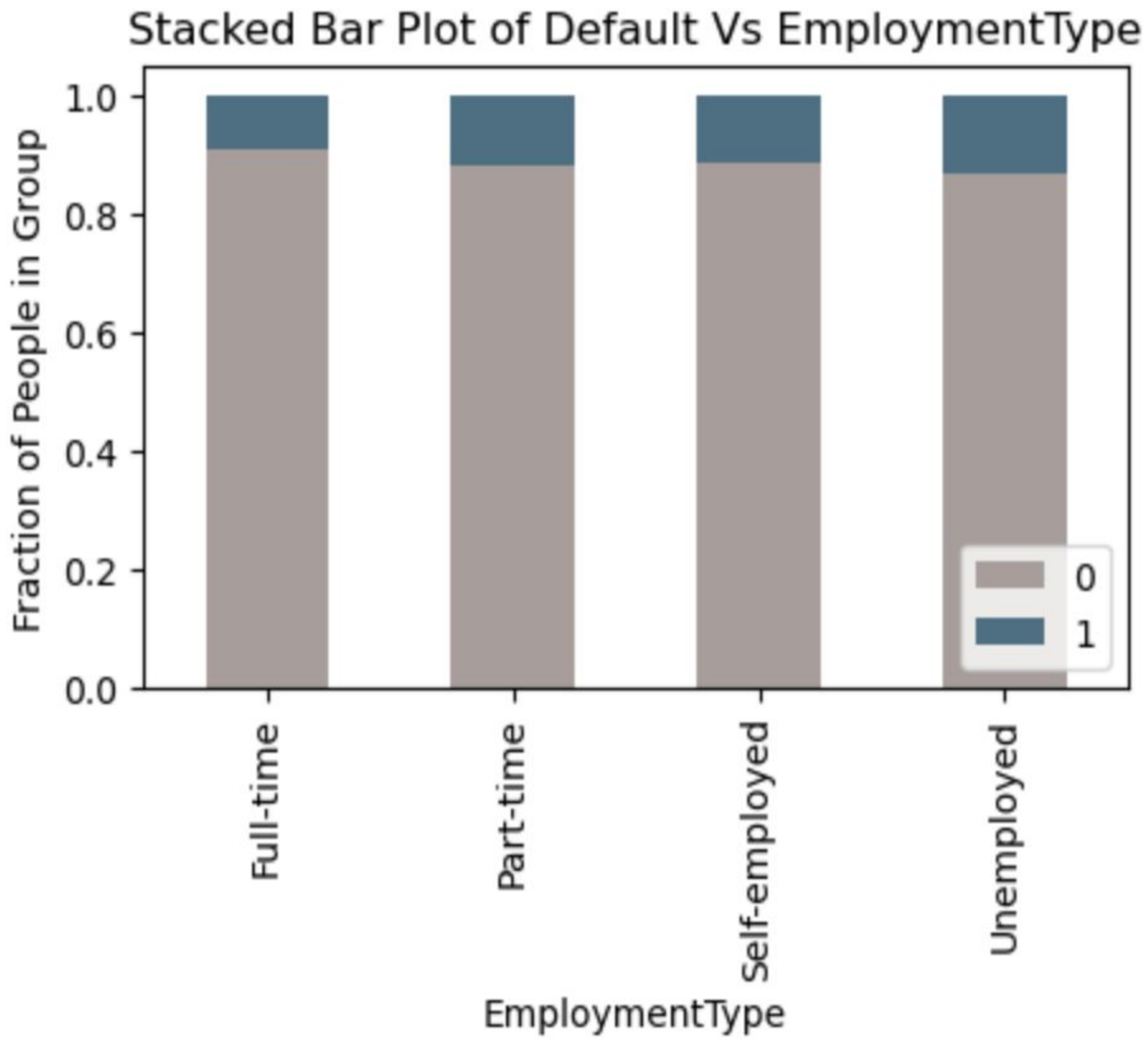
03

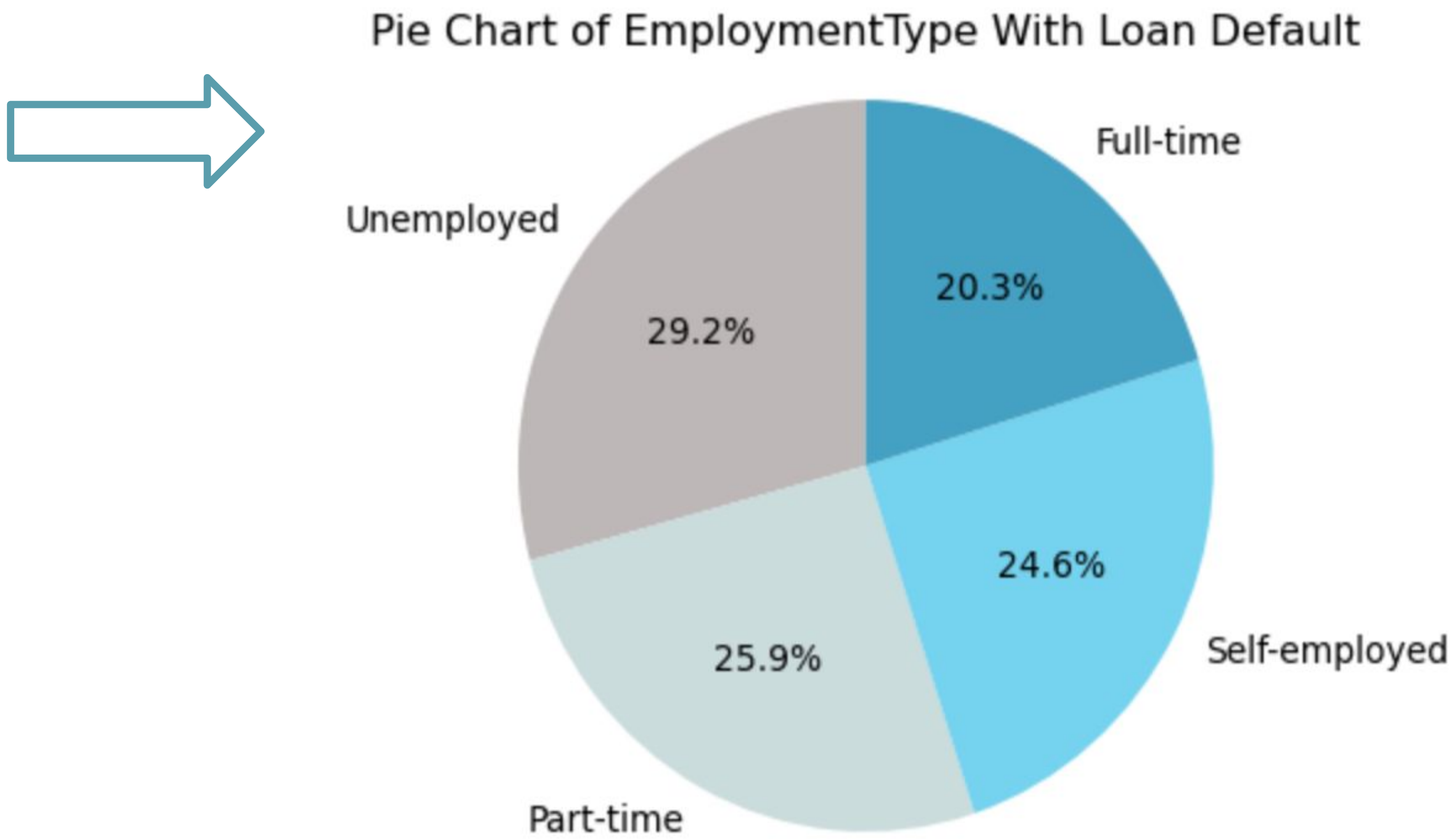
Continuous Features



There is no potential concerns for removing any high-correlated continuous features.

EDA Part II: Visualization of Column Pairs



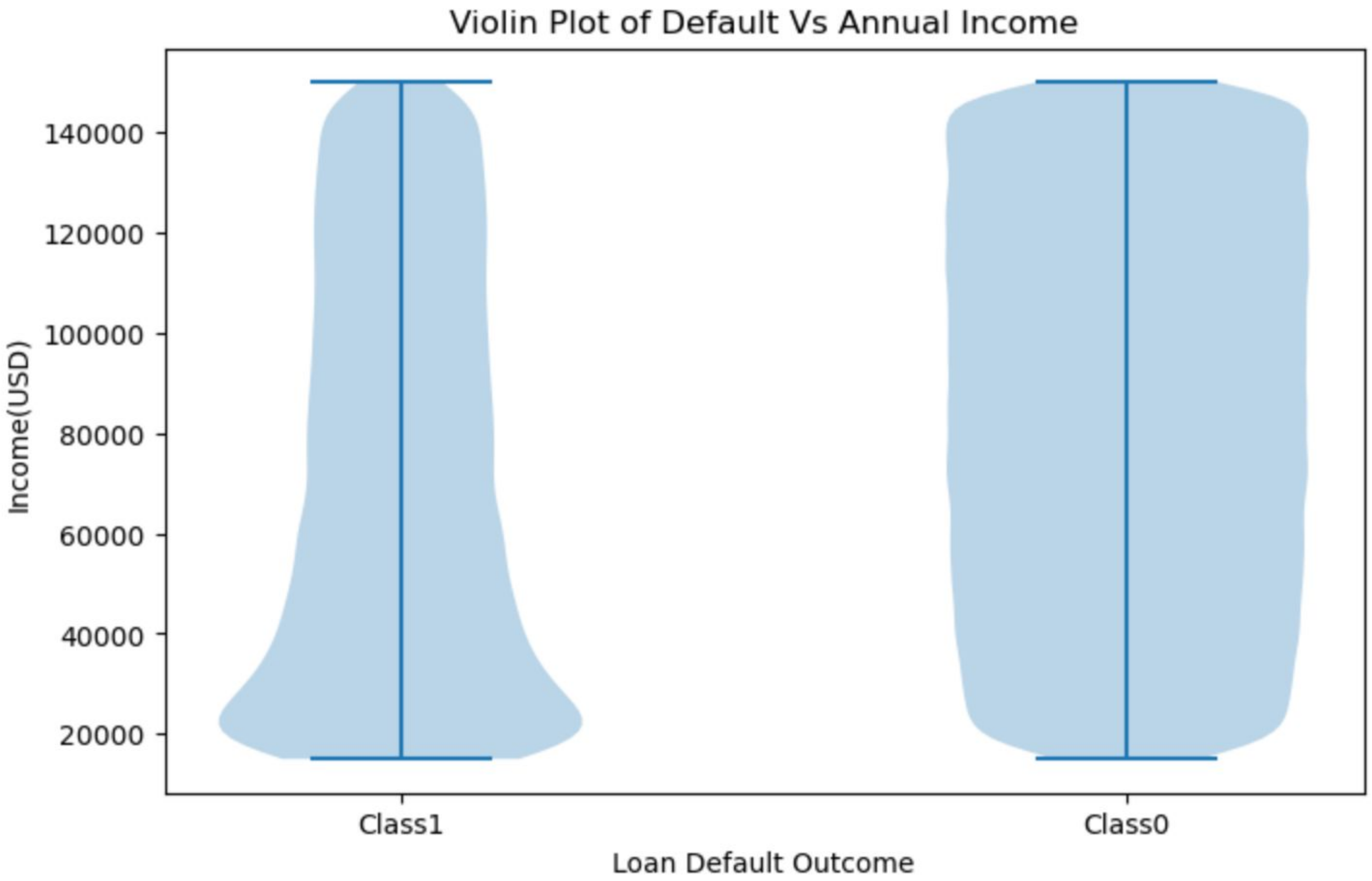


Relationship Between Loan Default and Employment Type:

Although there doesn't seem quite a difference, people who are employed full-time are more unlikely to default on their loans than people who are unemployed.

Relationship Between Loan Default and Annual Income:

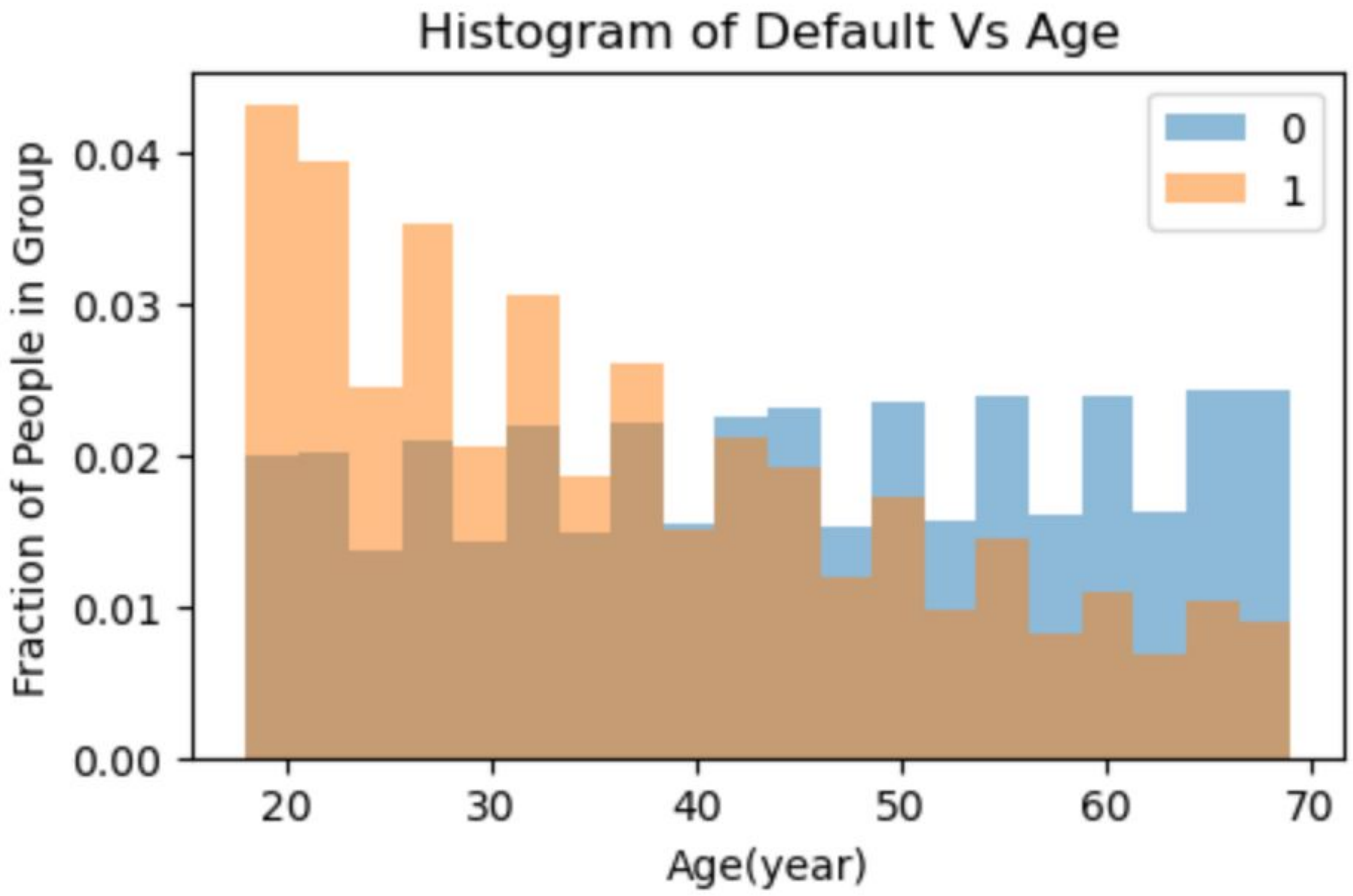
The first violin is obviously highly tailed, indicating that those with lower incomes are far more likely to encounter loan default.



Relationship Between Loan Default and Age:

The class1 distribution for the category-specific histogram is extremely skewed.

The likelihood of a loan default is four times higher for young individuals in their 20s than for older individuals in their 60s.



Missing Values&Group Structure:

Check before splitting and preprocessing:

Missing Value

Summary of Missing Values:

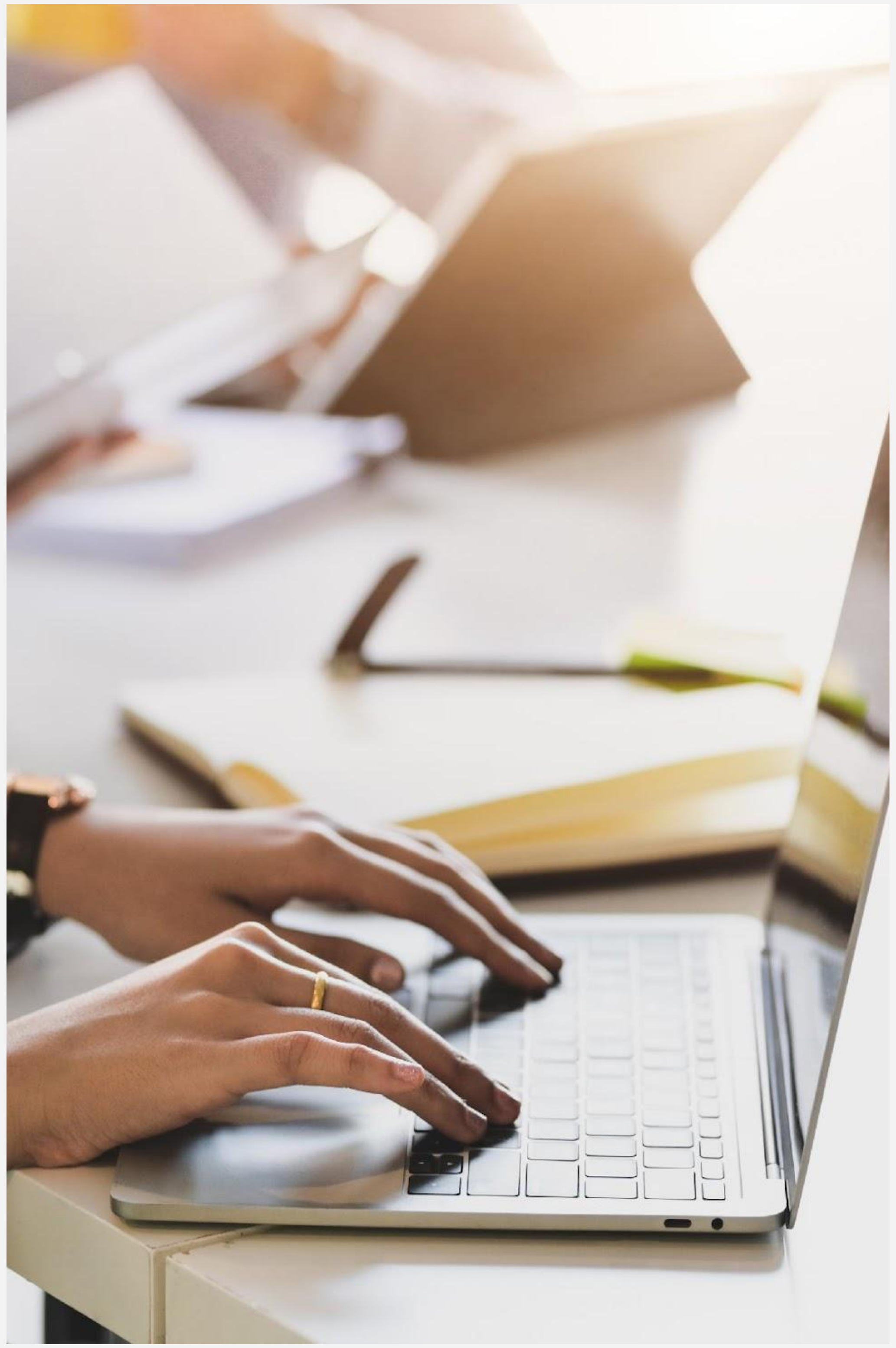
LoanID	0
Age	0
Income	0
LoanAmount	0
CreditScore	0
MonthsEmployed	0
NumCreditLines	0
InterestRate	0
LoanTerm	0
DTIRatio	0
Education	0
EmploymentType	0
MaritalStatus	0
HasMortgage	0
HasDependents	0
LoanPurpose	0
HasCoSigner	0
Default	0

IID Dataset

- No duplicate rows
- Drop column ‘LoanID’:
  a unique identifier represent different individuals borrowed loans.

Therefore, no group structure!

Splitting:



Stratify Splitting:

- Computationally efficient for this dataset with over 250k rows.
- Train-Validation-Test Ratio: 60-20-20

Stratified KFold:

- Generate 5 splits with same 60-20-20 train/validation/test ratio but more data will be trained.
- The most computationally intensive to train the models but would help in subsequent steps in a ML pipeline.

Output:

Shape	Class0	Class1	Total
Train	(135416,16)	(17792,16)	(153208,16)
Validation	(45139,16)	(5930,16)	(51069,16)
Test	(45139,16)	(5931,16)	(51070,16)

Preprocessing

One-Hot Encoding - - - **Ordinal Encoding** - - - **Standard Scaler**

6 Categorical Features:

- EmploymentType
- MaritalStatus
- LoanPurpose
- HasMortgage
- HasDependents
- HasCoSigner

1 Ordinal Feature:

- Education

  Order:

  - ❑ High School,
  - ❑ Bachelor's,
  - ❑ Master's,
  - ❑ PhD

9 Continuous Features:

- Age
- Income
- LoanAmount
- CreditScore
- MonthsEmployed
- 'NumCreditLines
- InterestRate'
- LoanTerm
- DTIRatio

Shape Before Preprocessing:

Train: (153208,16)
Validation: (51069,16)
Test: (51070,16)

Shape After Preprocessing:

Train: (153208,28)
Validation: (51069,28)
Test: (51070,28)

12 columns are added!

THANK YOU

10/23/24