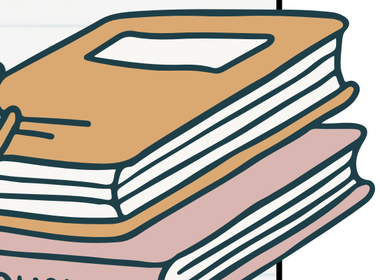
University of Ìnormation Technology
Data mining and applications - CS313

Loan approval prediction




Member

Member	Task
Nguyễn Đông Anh - 21520569	Slide (16%)
Huỳnh Nhật Hòa - 21520860	Orientation, feedback, experiment (18%)
Nguyễn Hà Anh Vũ - 21520531	EDA, Demo (18%)
Nguyễn Đinh Minh Chí - 21520648	Present (17%)
Trần Công Hải – 21520811	Report (16%)
Phạm Đức Hiếu - 21520856	Report, Slide (15%)



Overview

1. Introduction
2. Problem Setup
3. Explore Data Analysis
4. Method
5. Experiments
6. Conclusions

Introduction

Context

- Personal credit plays an important role in the economy.
- Despite that, risks of bad debt can have negative impacts on the economy.

Target:

- Building a model to predict whether a customer is likely to repay a loan, based on features such as income, credit history, and other related information.

Introduction

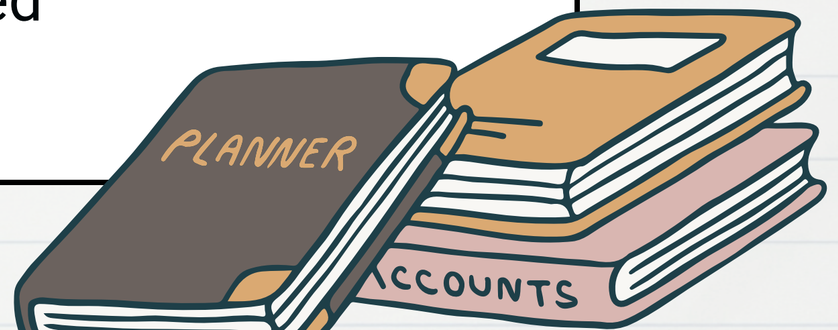
Motivation:

- Minimize credit risk.
- Optimize loan portfolio profitability.

Problem setup

1.Problem definition

- *Input*: Personal information about customers such as: income, credit history, current loans, age, ...
- *Output*: provide loan status prediction
  - Label 0: it means the customer is likely to be able to repay the loan -> loan approved
  - Label 1: it means the customer is likely to be unable to repay the loan -> loan not approved



Problem setup

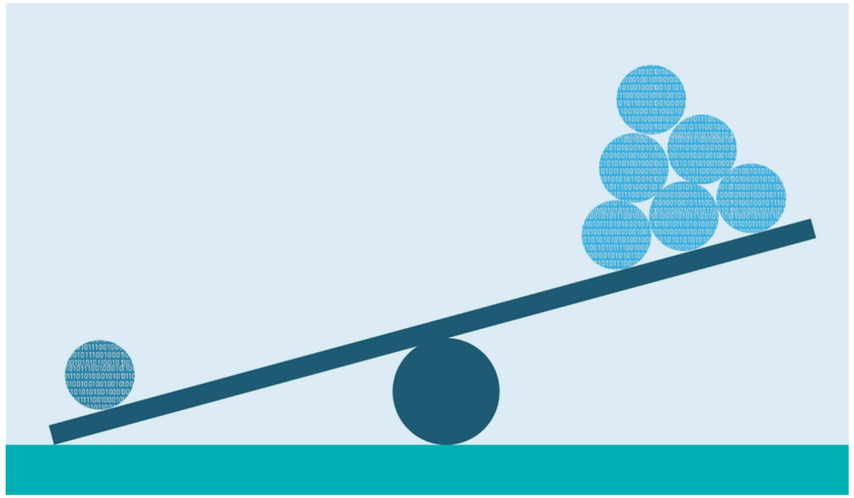
2.Data features:

- Source: Dataset train.csv và test.csv from Kaggle
- Samples and variable:
  - Samples: 58645
  - Number of features: 11 (7 numeric variables and 4 categorical variables)

Problem setup

2.Data features:

- *Challenge*:
  - Distribution: imbalanced target variable.
  - Multi types of variable.



Explore Data Analysis

View data :

train.csv

id	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length	loan_status
0	37	35000	RENT	0.0	EDUCATION	B	6000	11.49	0.17	N	14	0
1	22	56000	OWN	6.0	MEDICAL	C	4000	13.35	0.07	N	2	0
2	29	28800	OWN	8.0	PERSONAL	A	6000	8.90	0.21	N	10	0

...

test.csv

id	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
58645	23	69000	RENT	3.0	HOMEIMPROVEMENT	F	25000	15.76	0.36	N	2
58646	26	96000	MORTGAGE	6.0	PERSONAL	C	10000	12.68	0.10	Y	4
58647	26	30000	RENT	5.0	VENTURE	E	4000	17.19	0.13	Y	2

...

```
TARGET: loan_status
The number of feature: 11
==============================
The number of Numerical features: 7
Numerical features: ['person_age', 'person_income', 'person_emp_length', 'loan_amnt', 'loan_int_rate', 'loan_percent_income', 'cb_person_cred_hist_length']
==============================
The number of Categorical features: 4
Categorical features: ['person_home_ownership', 'loan_intent', 'loan_grade', 'cb_person_default_on_file']
```

Explore Data Analysis

View data :

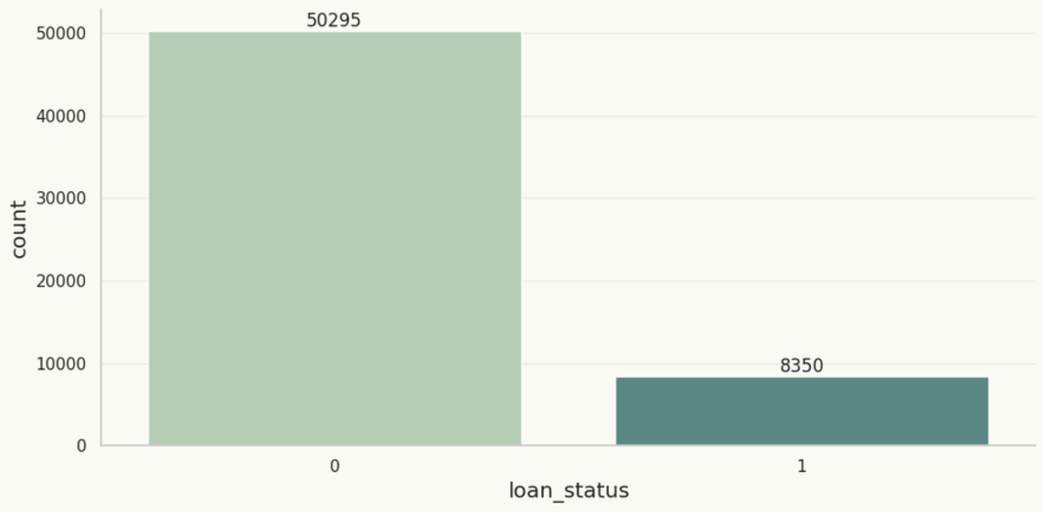
Missing value

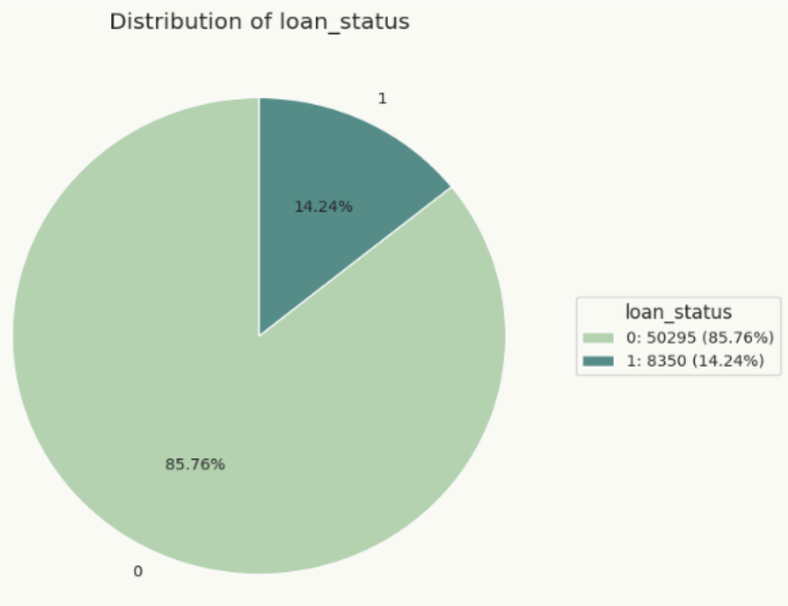
column	train.csv	test.csv
id	0	0
person_age	0	0
person_income	0	0
person_home_ownership	0	0
person_emp_length	0	0
loan_intent	0	0
loan_grade	0	0
loan_amnt	0	0
loan_int_rate	0	0
loan_percent_income	0	0
cb_person_default_on_file	0	0
cb_person_cred_hist_lengt	0	0
loan_status	0	X

	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_percent_income	cb_person_cred_hist_length
count	58645.000000	5.864500e+04	58645.000000	58645.000000	58645.000000	58645.000000	58645.000000
mean	27.550857	6.404617e+04	4.701015	9217.556518	10.677874	0.159238	5.813556
std	6.033216	3.793111e+04	3.959784	5563.807384	3.034697	0.091692	4.029196
min	20.000000	4.200000e+03	0.000000	500.000000	5.420000	0.000000	2.000000
25%	23.000000	4.200000e+04	2.000000	5000.000000	7.880000	0.090000	3.000000
50%	26.000000	5.800000e+04	4.000000	8000.000000	10.750000	0.140000	4.000000
75%	30.000000	7.560000e+04	7.000000	12000.000000	12.990000	0.210000	8.000000
max	123.000000	1.900000e+06	123.000000	35000.000000	23.220000	0.830000	30.000000

Explore Data Analysis

The distribution of TARGET





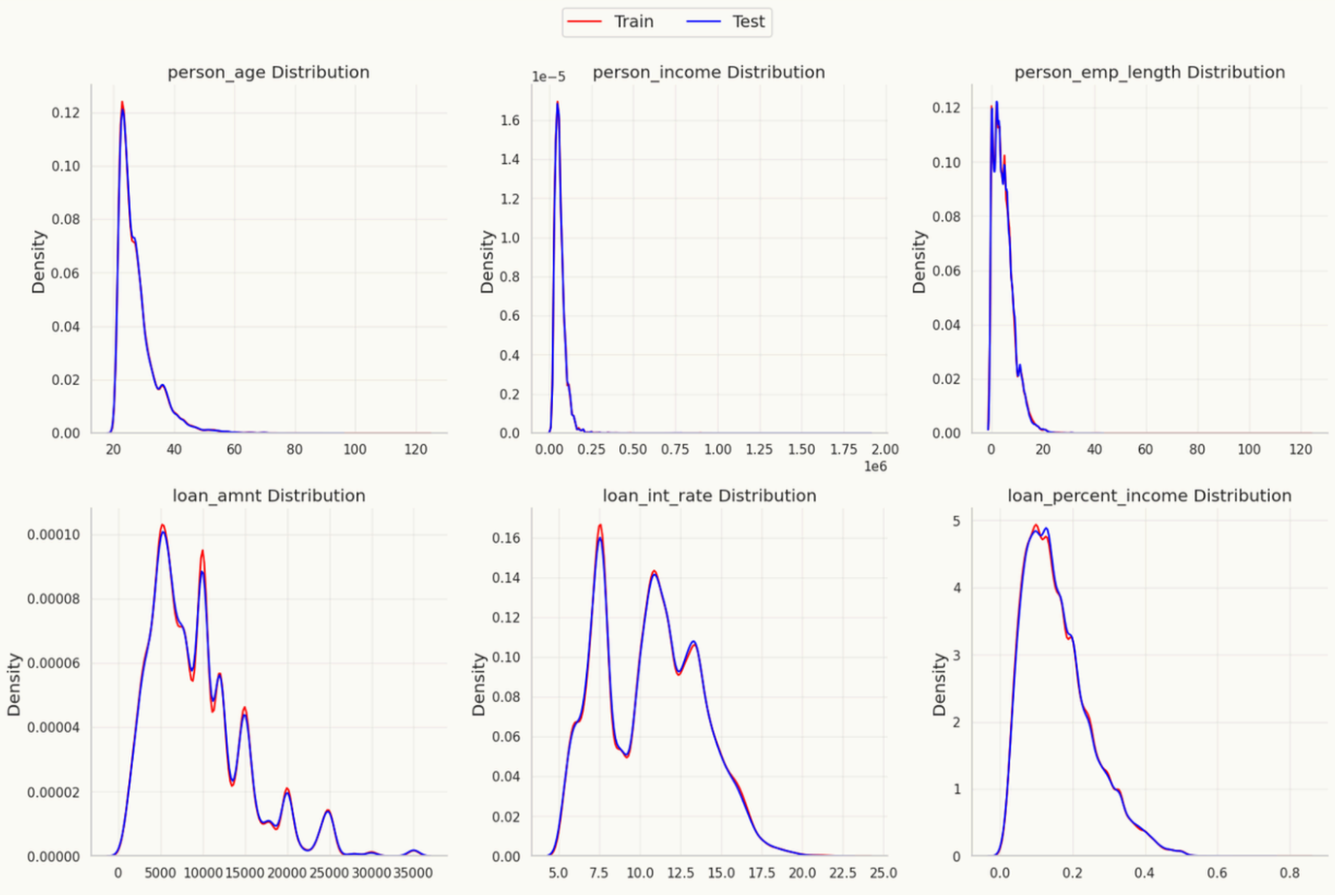
IR = 6.02 > 5
ENTROPY: 0.4093
CHI-SQUARE STATISTIC: 30000.5631,
P-VALUE: 0.0000

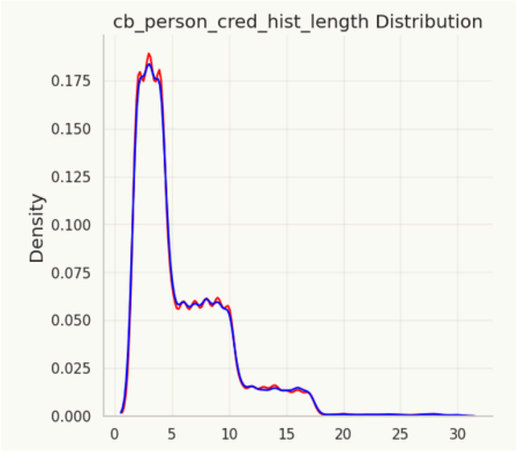
=> Serious imbalance

The distribution of 'loan_status' shows an imbalance, so we use models like CatBoost, XGBoost, ...
Beside that, gradient Boosting-based algorithms are effective in handling imbalanced datasets.

Explore Data Analysis

Distribution of Numeric features on train and test

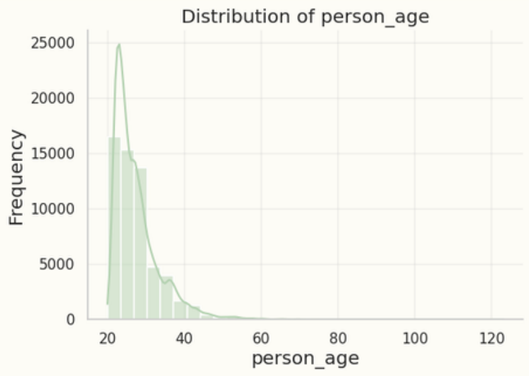


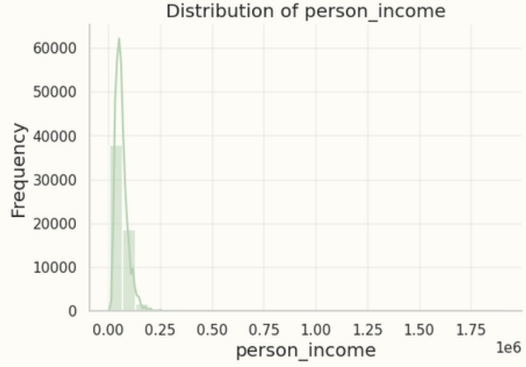


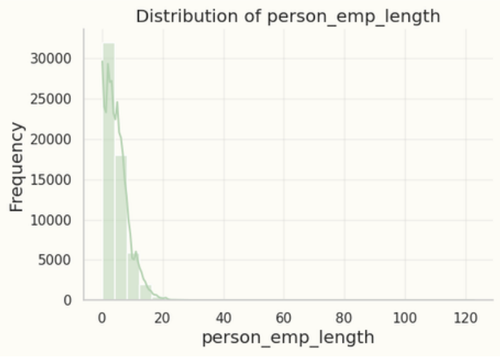
KERNEL DENSITY ESTIMATION (KDE) PLOT

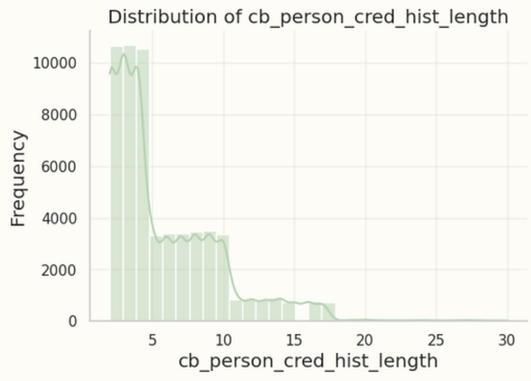
Explore Data Analysis

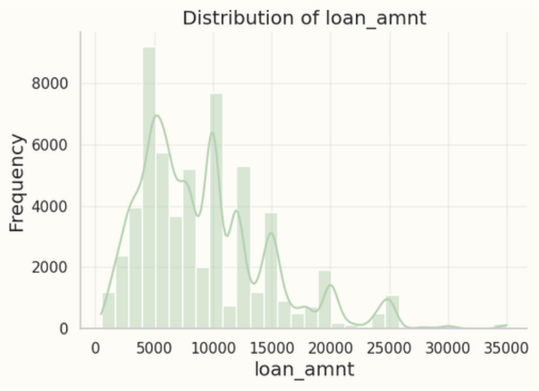
Distribution of Numeric features

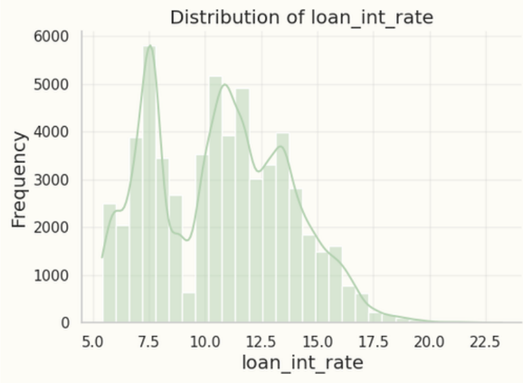


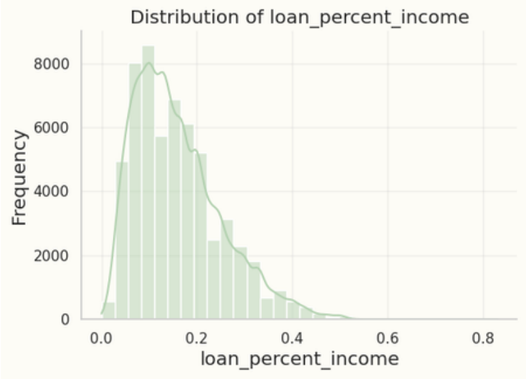






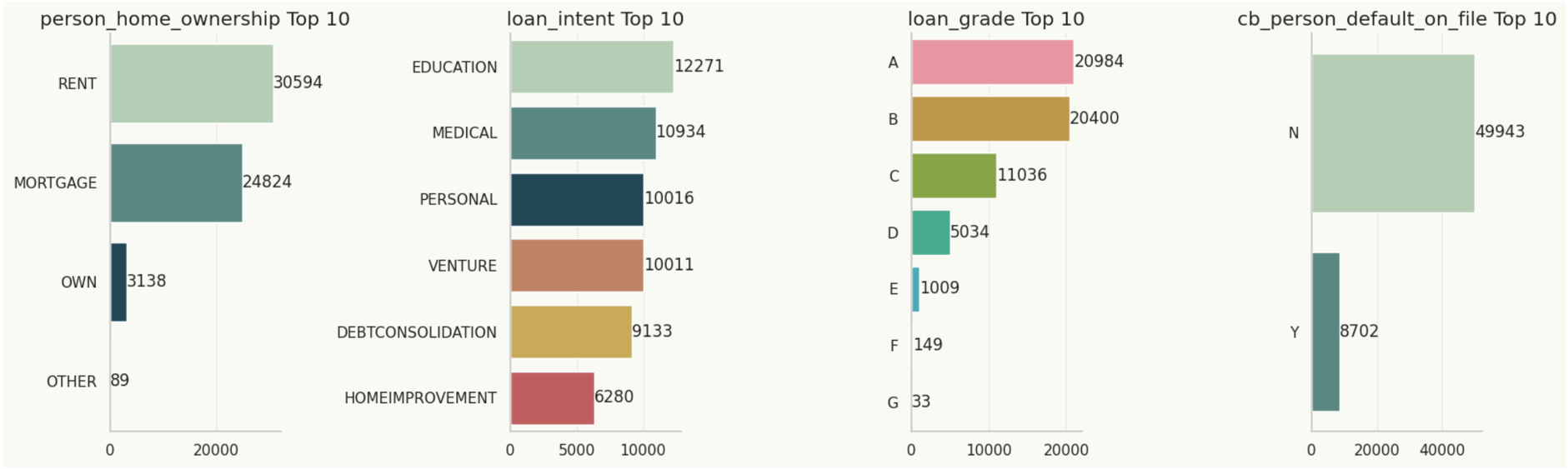






Explore Data Analysis

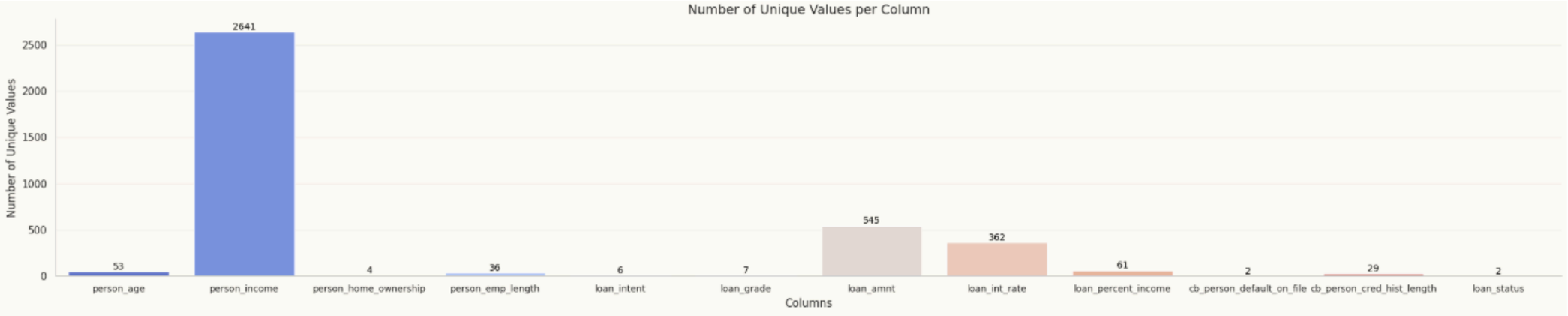
Distribution of categorical features



Explore Data Analysis

Unique values

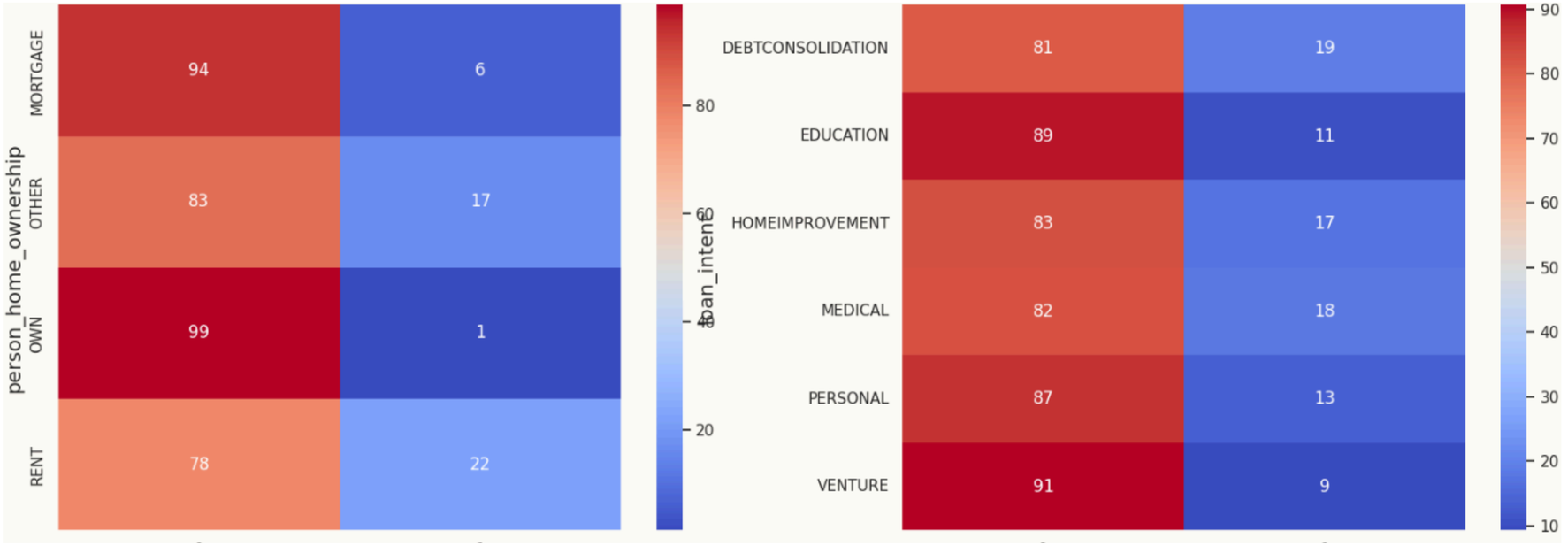
Chart below performs number of unique values of each columm.



Explore Data Analysis

Visualize data :

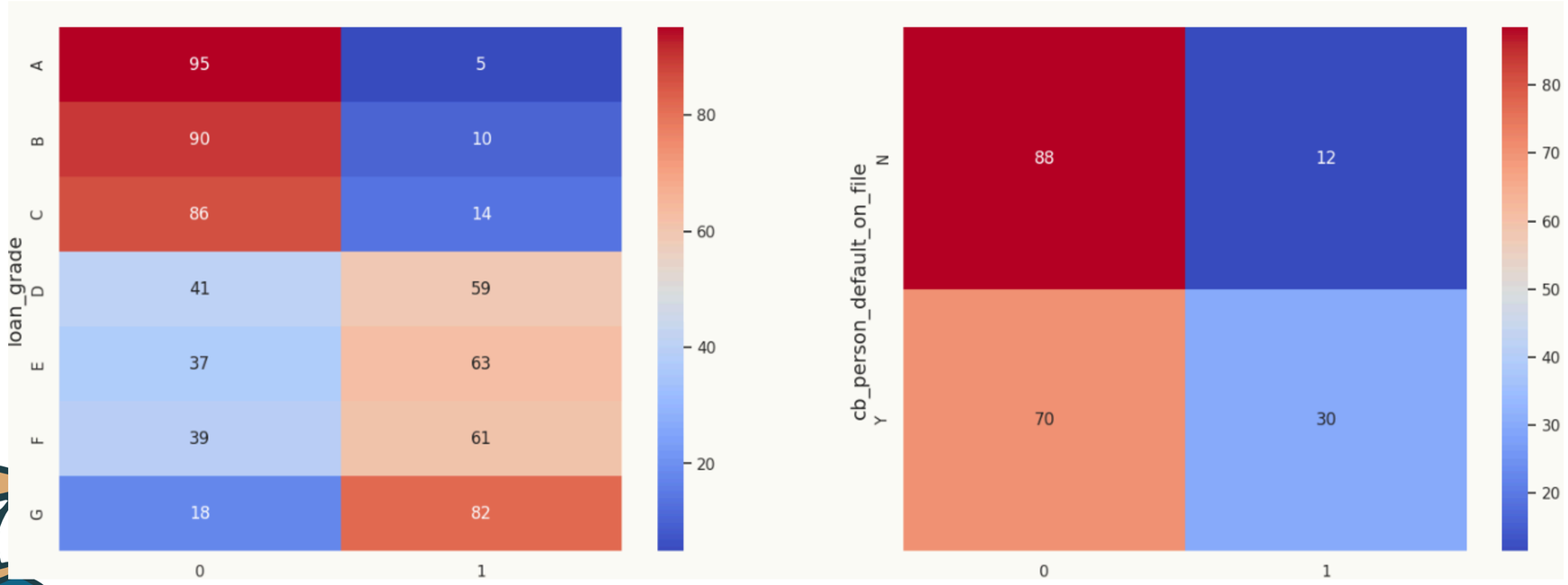
Some chart here show the dependency between loan status and a couple of unique value.



Explore Data Analysis

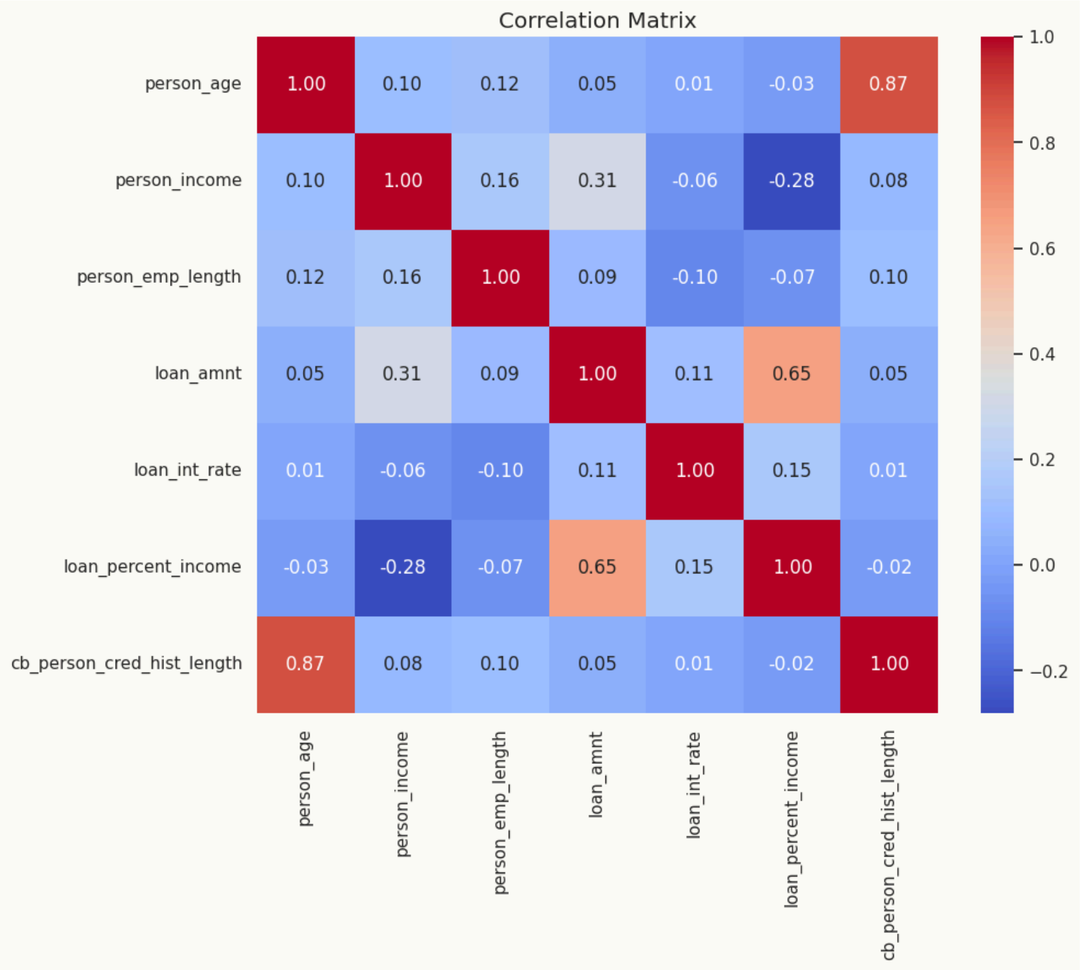
Visualize data :

Some chart here show the dependency between loan status and a couple of unique value.



Explore Data Analysis

Correlation matrix:



Method

1. Data processing procedures

Data cleaning:

- Fill missing values using median (numeric variables) and mode (categorical variables). However, this is not necessary because there are no missing values.
- Remove samples with extreme outliers: also not necessary because we will use meta-models such as XGBoost, CatBoost, and LightGBM, which are robust to outliers.

Feature Engineering

- Create new variables, such as debt_to_income_ratio: We tried this, but it didn't improve the model's performance significantly.
- Transform numeric variables into categorical variables (while keeping the original variables): This approach proved to be more effective.

Method

1. Data processing procedures

Encoding

- One-hot: Used for variables with few distinct values, such as **'loan_intent'**.
- Ordinal: Used for variables with many distinct values that have a natural order.

Scaling

- No need

Handling Imbalanced Data

- Using cross validation.
- Using models that handle imbalanced data well.

Method

2. Model Used & Evaluation Metrics

XGBoost

High performance with tabular data, and has mechanisms for handling class imbalance.

CatBoost

Automatically supports categorical data with minimal preprocessing required.

LightGBM

Fast and efficient with large datasets.

AUC-ROC

Measures the ability of a model to distinguish between classes.

EXPERIMENTS

Table of results

	Private	Public
XgBoost(1)	0.95679	0.96096
Catboost(2)	0.95717	0.95728
LightGBM(3)	0.95862	0.95768
Voting Classifier(1+2+3)	0.95954	0.96028
Logistic Regression	0.84918	0.84229

Conclusion

1.Result summary

In the loan approval prediction task, based on features like financial status, loan purpose,... , models such as Boosting (XGBoost, LightGBM, CatBoost) and Voting Ensembles have demonstrated strong potential and effectiveness in predicting approval outcomes.

2.Further development

- Try other method to improve predictions

- Find out more suitable parameters .for model such as grid search, random search, optune ...

- A more details dataset will provide better predictions.

Thank's For Watching