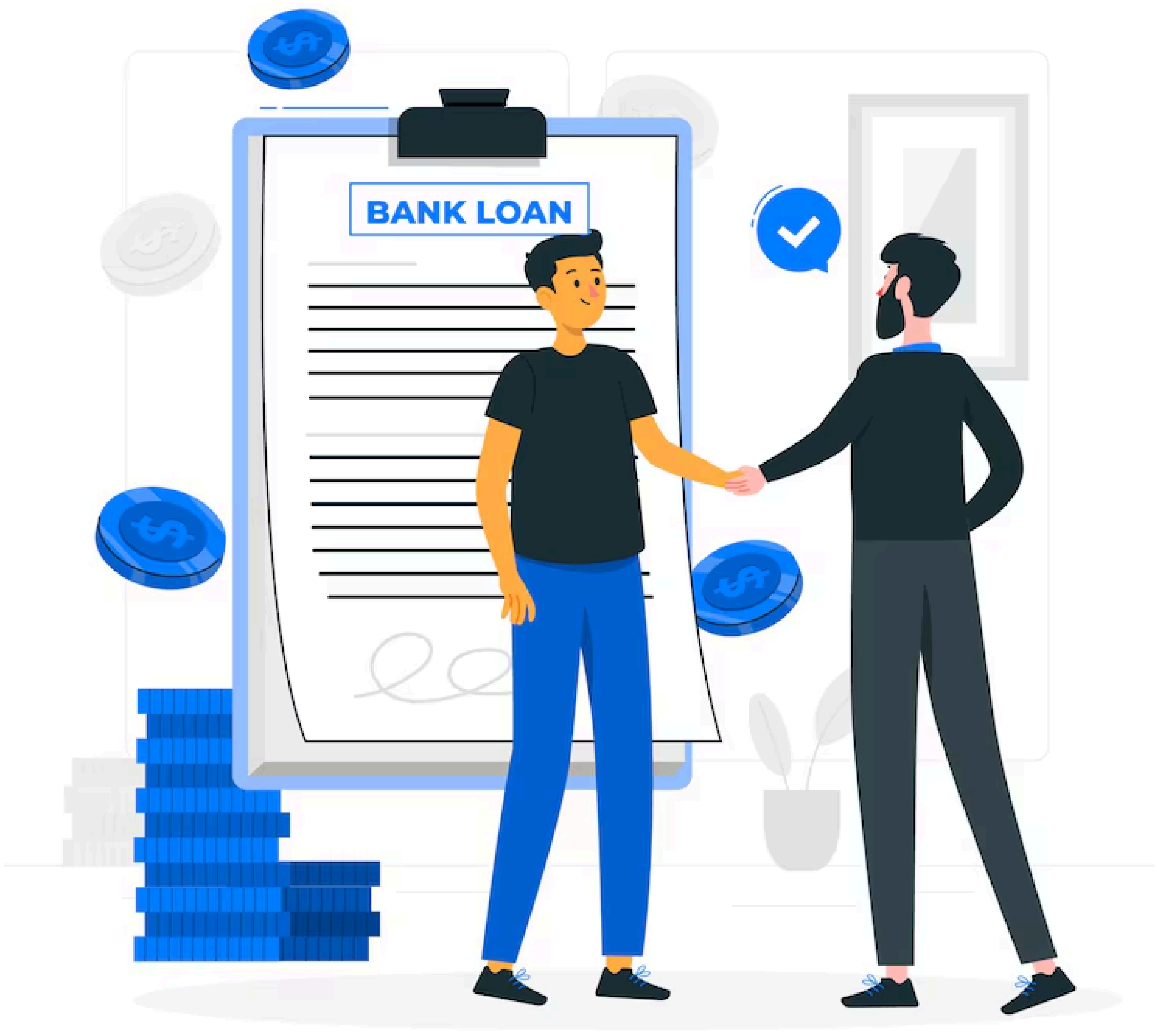


LOAN APPROVAL PREDICTION

23521570 - Huỳnh Việt Tiến
23520123 Nguyễn Minh Bảo
23520133 - Phạm Phú Bảo
23521143 - Nguyễn Công Phát

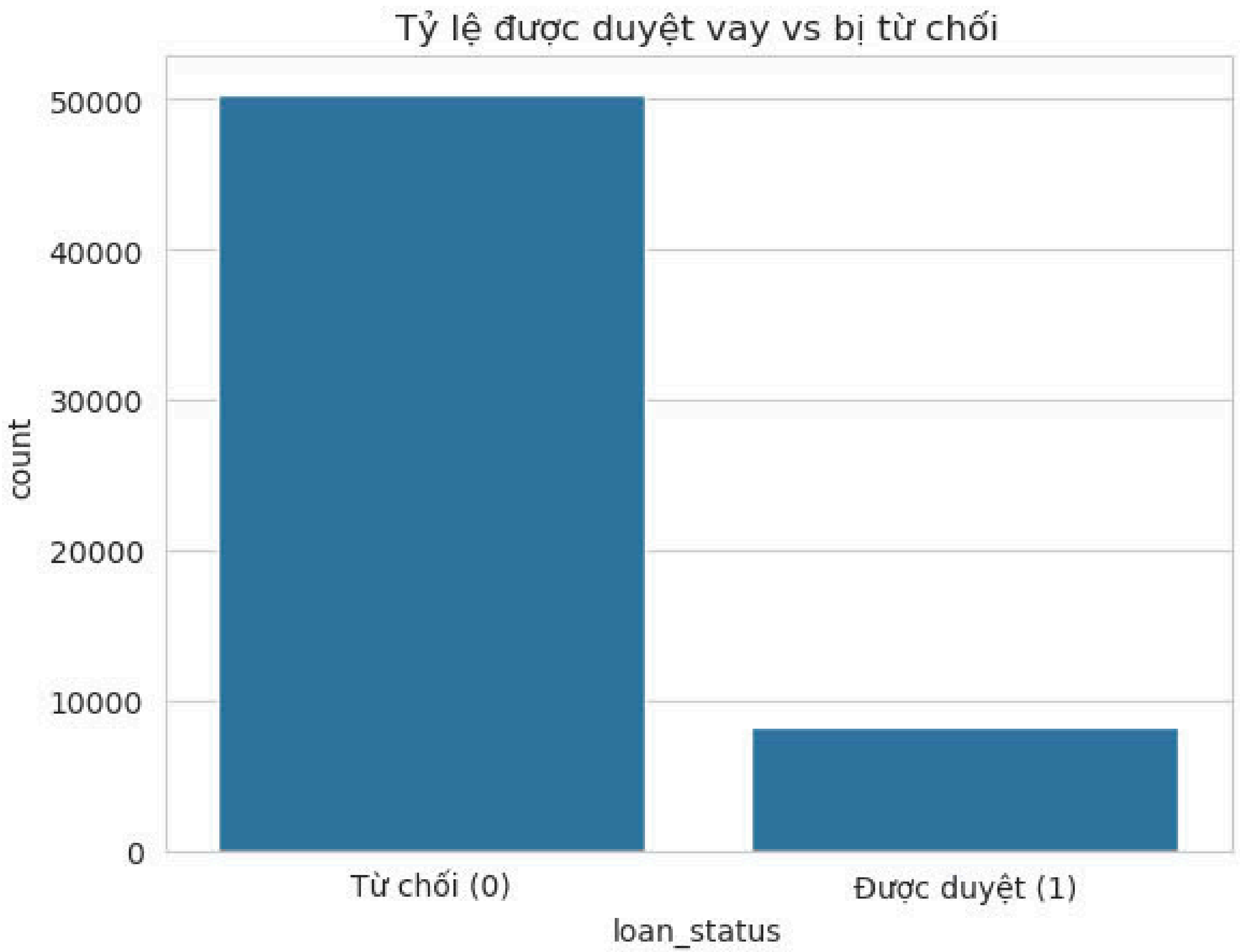
Table of content

EXPLORATORY DATA ANALYSIS

Loan Approval Prediction

Introduction

- 58644 Samples
- 7 Numerical Features
- 4 Categorical Features
- Target Feature: loan_status(0,1)



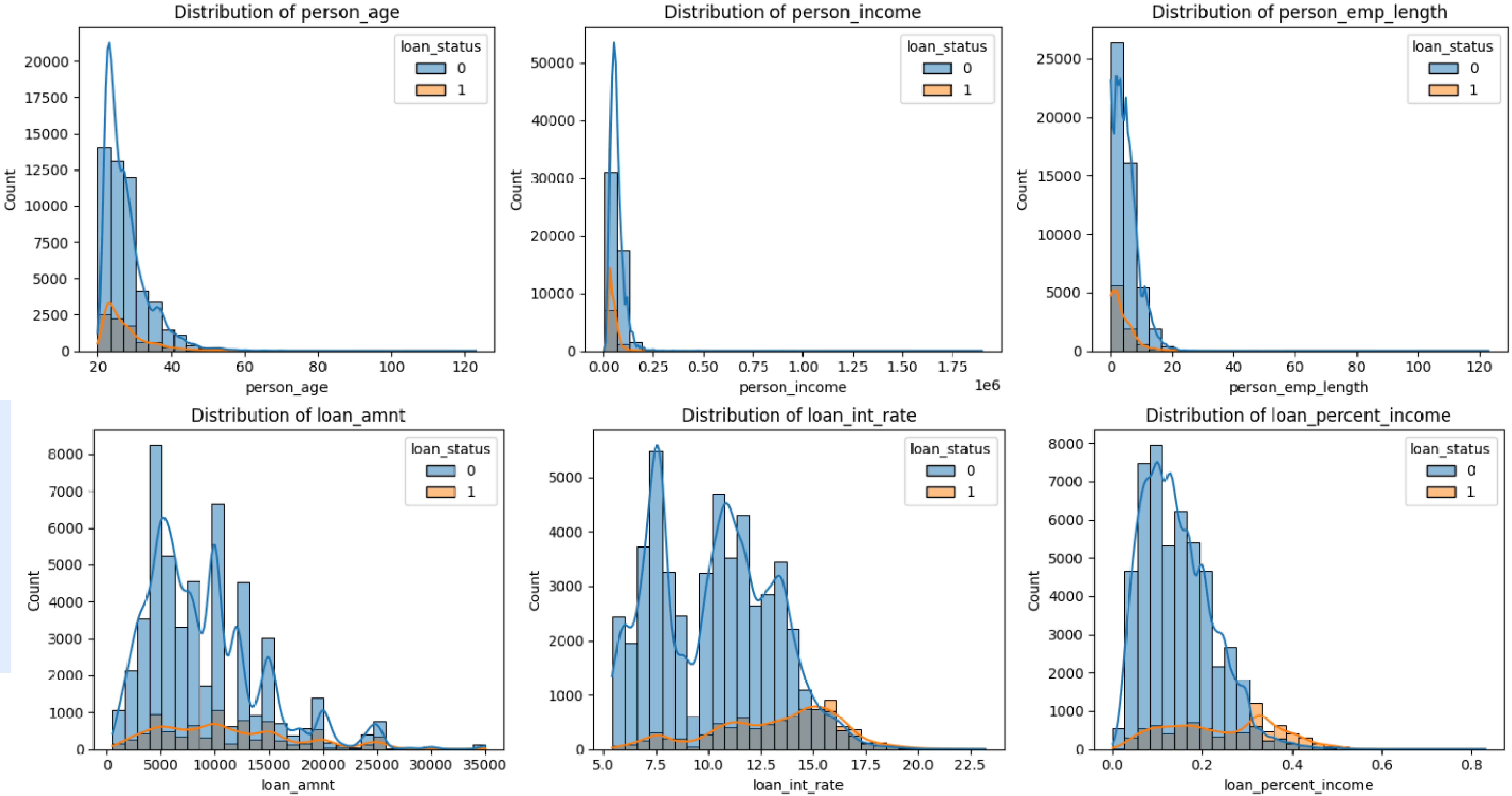
Explain Numerical Features:

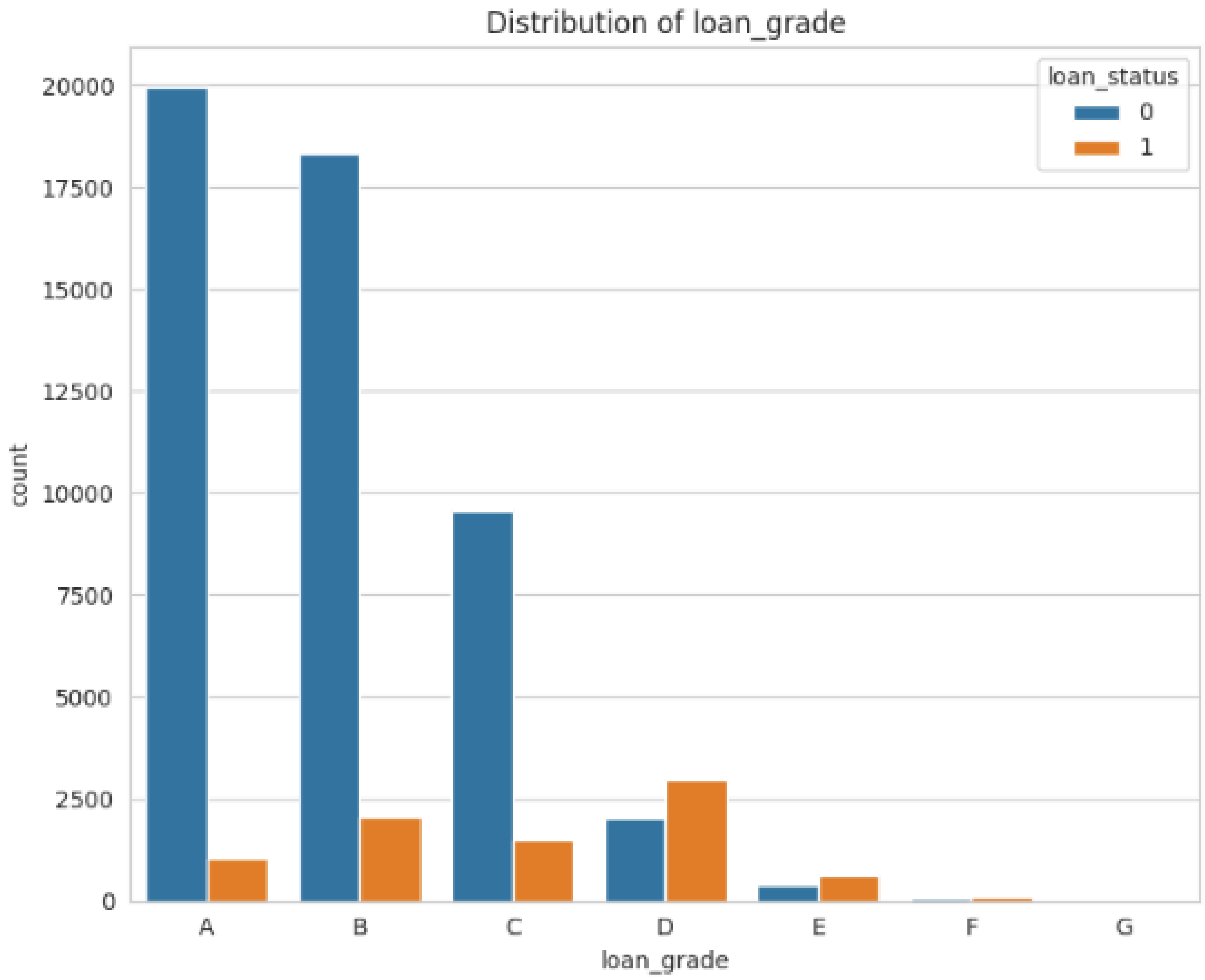
- **person_age:** Age of the applicant.
- **person_income:** Income of the applicant.
- **loan_amnt:** Loan amount(USD).
- **loan_int_rate:** Loan interest rate(USD).
- **loan_percent_income:** Percentage of income allocated for the loan.
- **cb_person_cred_hist_length:** Length of the applicant's credit history.
- **person_emp_length:** Working time(year).

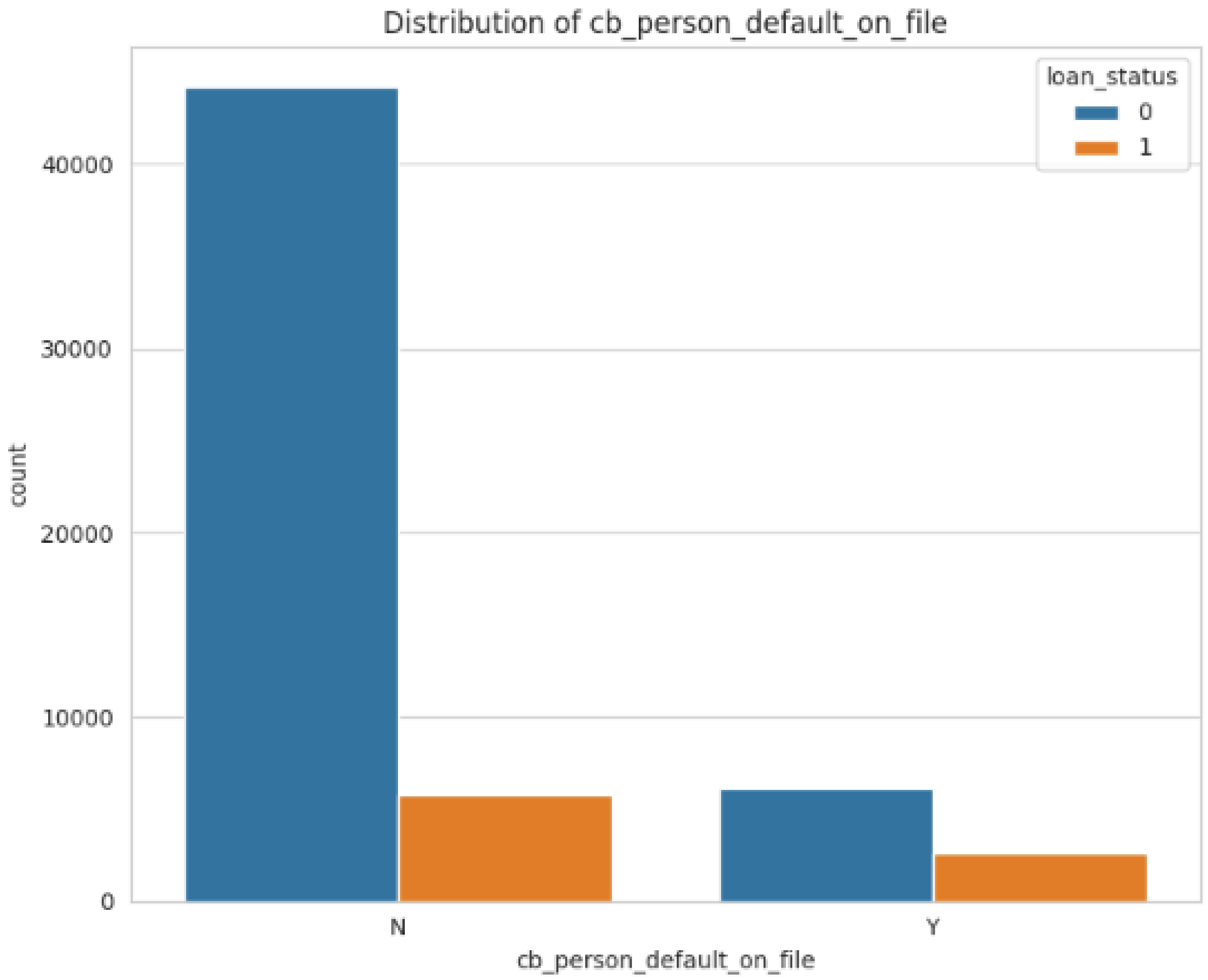
Explain categorial Features

- **person_home_ownership:** Home ownership status (RENT, OWN, MORTGAGE, OTHER).
- **loan_intent:** Purpose of the loan (EDUCATION, MEDICAL, VENTURE, PERSONAL, DEBTCONSOLIDATION)
- **loan_grade:** Loan rating/grade (A, B, C, D, E).
- **cb_person_default_on_file:** Credit default history (Y/N).

Bivariate Analysis




Distribution of loan_grade
loan_status
0
1
20000
17500
15000
12500
10000
7500
5000
2500
0
count
A
B
C
D
E
F
G
loan_grade

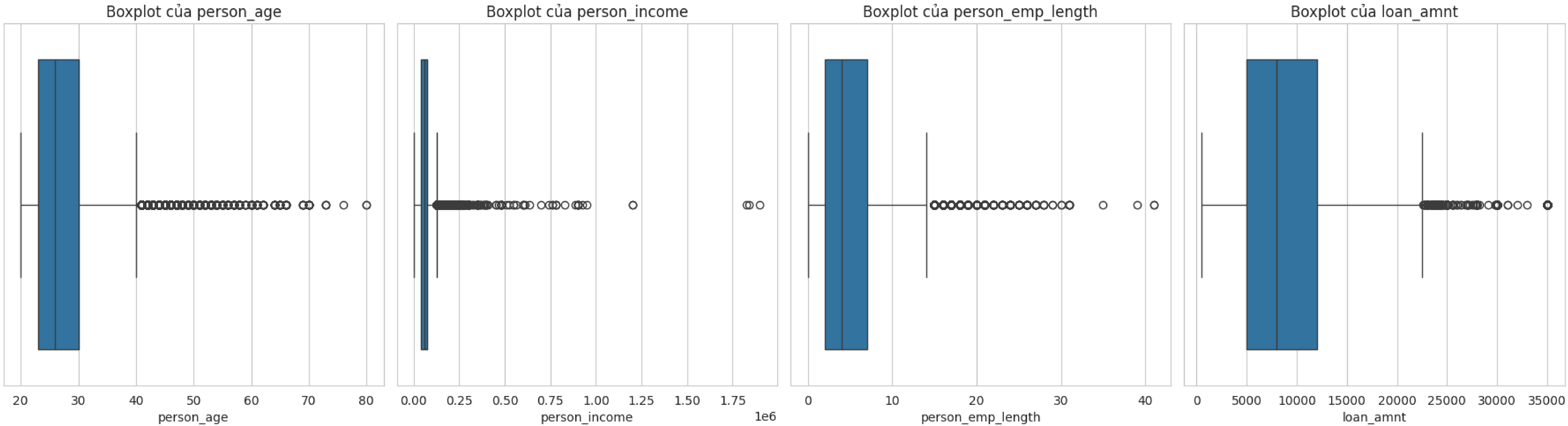

Distribution of cb_person_default_on_file
loan_status
0
1
40000
30000
20000
10000
0
count
N
Y
cb_person_default_on_file

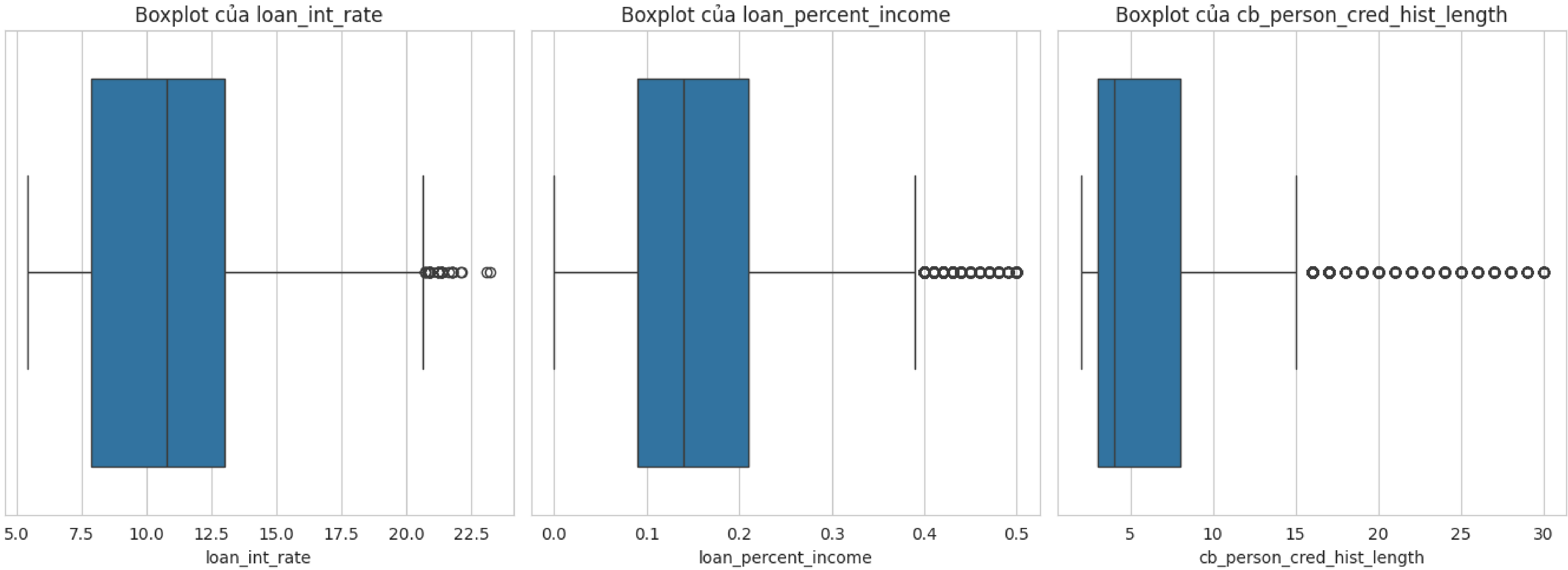
DATA PREPROCESSING

NULL

After applying techniques to detect NULL values, the result shows that there is no missing data.

	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
0	58645	23	69000	RENT	3.0	HOMEIMPROVEMENT	F	25000	15.76	0.36	N	2
1	58646	26	96000	MORTGAGE	6.0	PERSONAL	C	10000	12.68	0.10	Y	4
2	58647	26	30000	RENT	5.0	VENTURE	E	4000	17.19	0.13	Y	2
3	58648	33	50000	RENT	4.0	DEBTCONSOLIDATION	A	7000	8.90	0.14	N	7
4	58649	26	102000	MORTGAGE	8.0	HOMEIMPROVEMENT	D	15000	16.32	0.15	Y	4
5	58650	23	66000	RENT	5.0	EDUCATION	D	22000	14.09	0.33	N	2
6	58651	26	75000	OWN	10.0	PERSONAL	B	8000	10.62	0.11	N	4
7	58652	23	55000	MORTGAGE	6.0	PERSONAL	A	6250	6.76	0.12	N	2
8	58653	32	29124	RENT	0.0	PERSONAL	C	7200	13.11	0.26	Y	6
9	58654	22	90000	RENT	4.0	DEBTCONSOLIDATION	C	10000	13.49	0.11	Y	3


Boxplot của person_age
20 30 40 50 60 70 80
person_age
Boxplot của person_income
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75
1e6
person_income
Boxplot của person_emp_length
0 10 20 30 40
person_emp_length
Boxplot của loan_amnt
0 5000 10000 15000 20000 25000 30000 35000
loan_amnt


Boxplot của loan_int_rate
5.0 7.5 10.0 12.5 15.0 17.5 20.0 22.5
loan_int_rate
Boxplot của loan_percent_income
0.0 0.1 0.2 0.3 0.4 0.5
loan_percent_income
Boxplot của cb_person_cred_hist_length
5 10 15 20 25 30
cb_person_cred_hist_length

OUTLIER

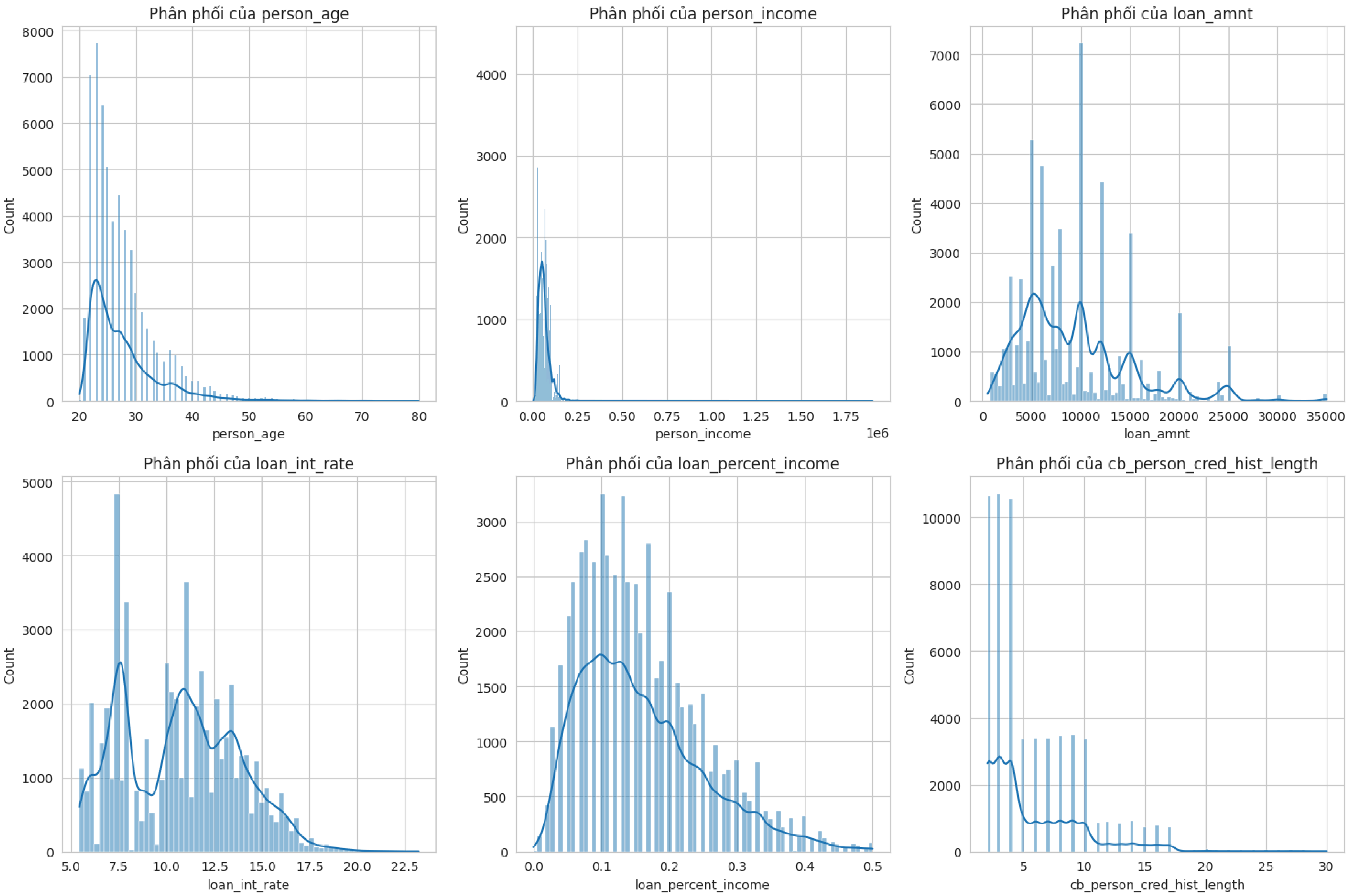
Encoding

Categorial Features

- **person_home_ownership:** Home ownership status (RENT, OWN, MORTGAGE, OTHER). **(Nominal)**
- **loan_intent:** Purpose of the loan (EDUCATION, MEDICAL, VENTURE, PERSONAL, DEBTCONSOLIDATION) **(Nominal)**
- **loan_grade:** Loan rating/grade (A, B, C, D, E). **(Ordinal)**
- **cb_person_default_on_file:** Credit default history (Y/N). **(Binary)**

	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	0	37	35000	RENT	0.0	EDUCATION	B	6000	11.49	0.17	N	14	0
1	1	22	56000	OWN	6.0	MEDICAL	C	4000	13.35	0.07	N	2	0
2	2	29	28800	OWN	8.0	PERSONAL	A	6000	8.90	0.21	N	10	0
3	3	30	70000	RENT	14.0	VENTURE	B	12000	11.11	0.17	N	5	0
4	4	22	60000	RENT	2.0	MEDICAL	A	6000	6.92	0.10	N	3	0
5	5	27	45000	RENT	2.0	VENTURE	A	9000	8.94	0.20	N	5	0
6	6	25	45000	MORTGAGE	9.0	EDUCATION	A	12000	6.54	0.27	N	3	0
7	7	21	20000	RENT	0.0	PERSONAL	C	2500	13.49	0.13	Y	3	0
8	8	37	69600	RENT	11.0	EDUCATION	D	5000	14.84	0.07	Y	11	0
9	9	35	110000	MORTGAGE	0.0	DEBTCONSOLIDATION	C	15000	12.98	0.14	Y	6	0

Data Scaling



→ **MinMaxScaler**

FEATURE ENIGNEERING

"Coming up with features is difficult, time-consuming, requires expert knowledge"
Andrew Ng

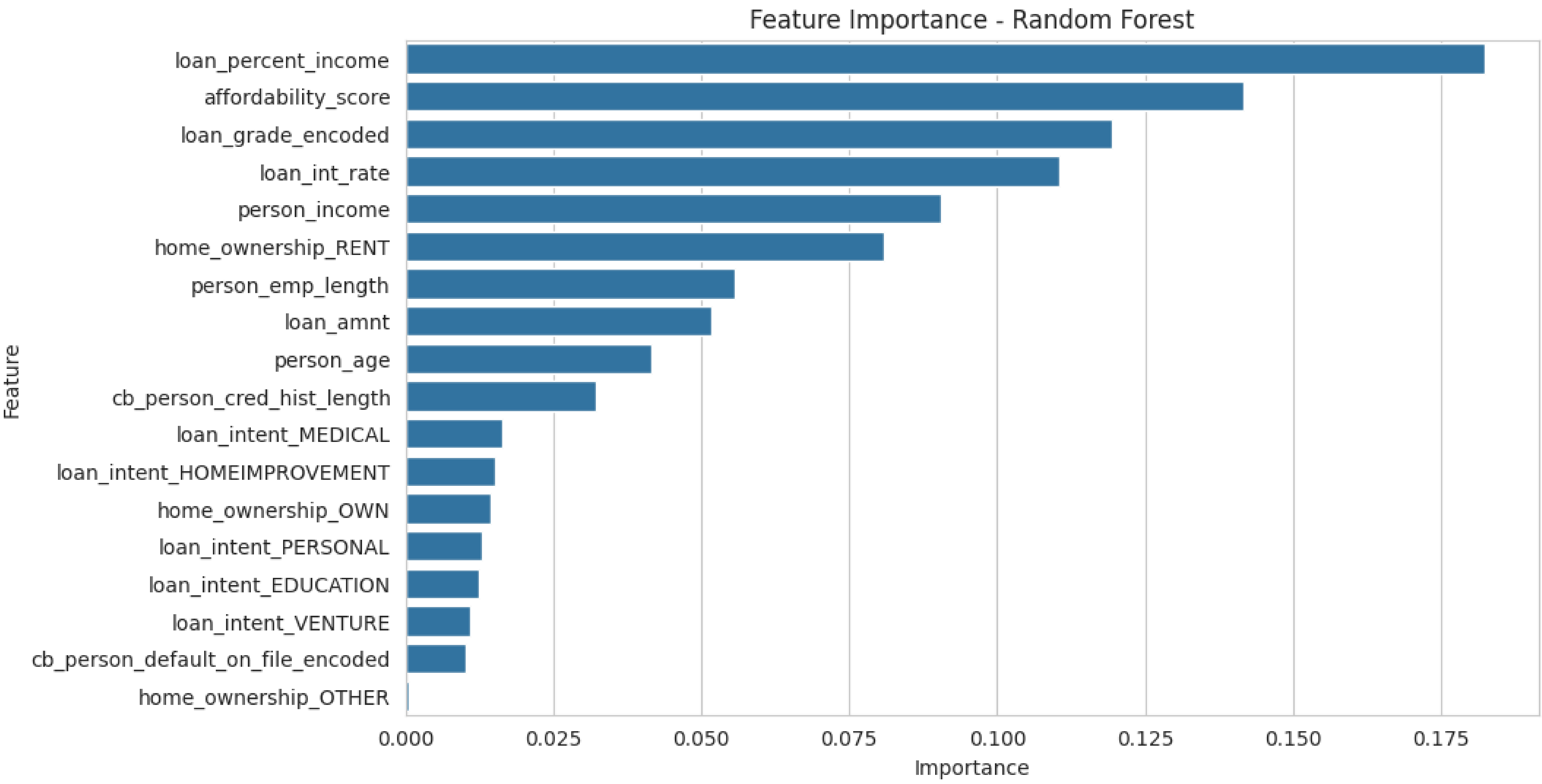
XGBOOST CLASSIFICATION FEATURE IMPORTANCE

$$\text{affordability\_score} = \frac{\text{person\_income}}{\left(\text{loan\_amnt} \times \frac{\text{loan\_int\_rate}}{100}\right)}$$

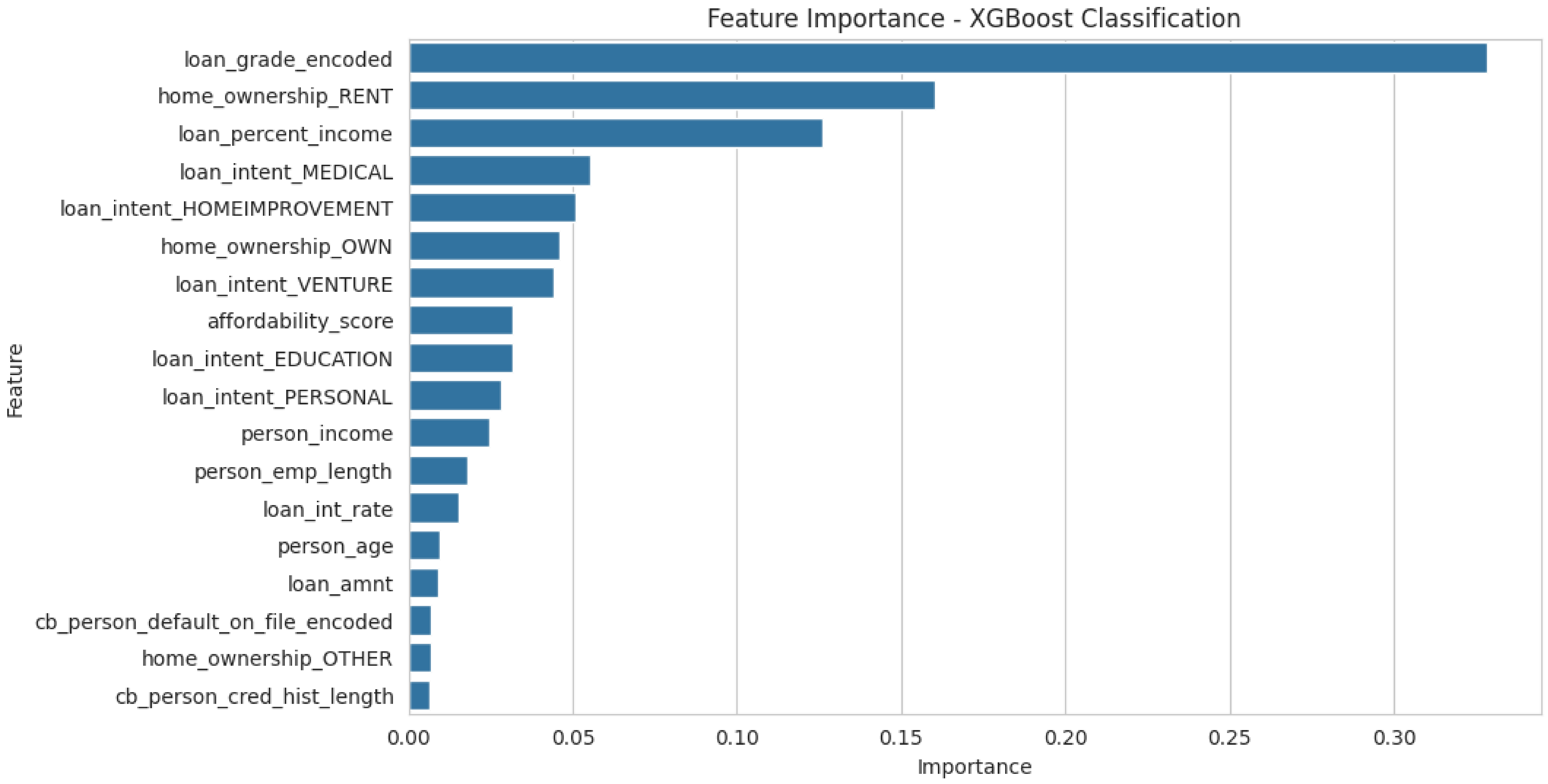
- Reflecting the actual repayment capacity of the borrower, including interest rates.
- Combining three important factors: income, loan amount, and interest rate
- Helping the model understand the financial stress level faced by the borrower.
- Distinguishing cases of high income but risk due to large loans/high interest rates.

→ Improving the model's ability to learn complex relationships, increasing prediction accuracy.

RANDOM FOREST FEATURE IMPORTANCE



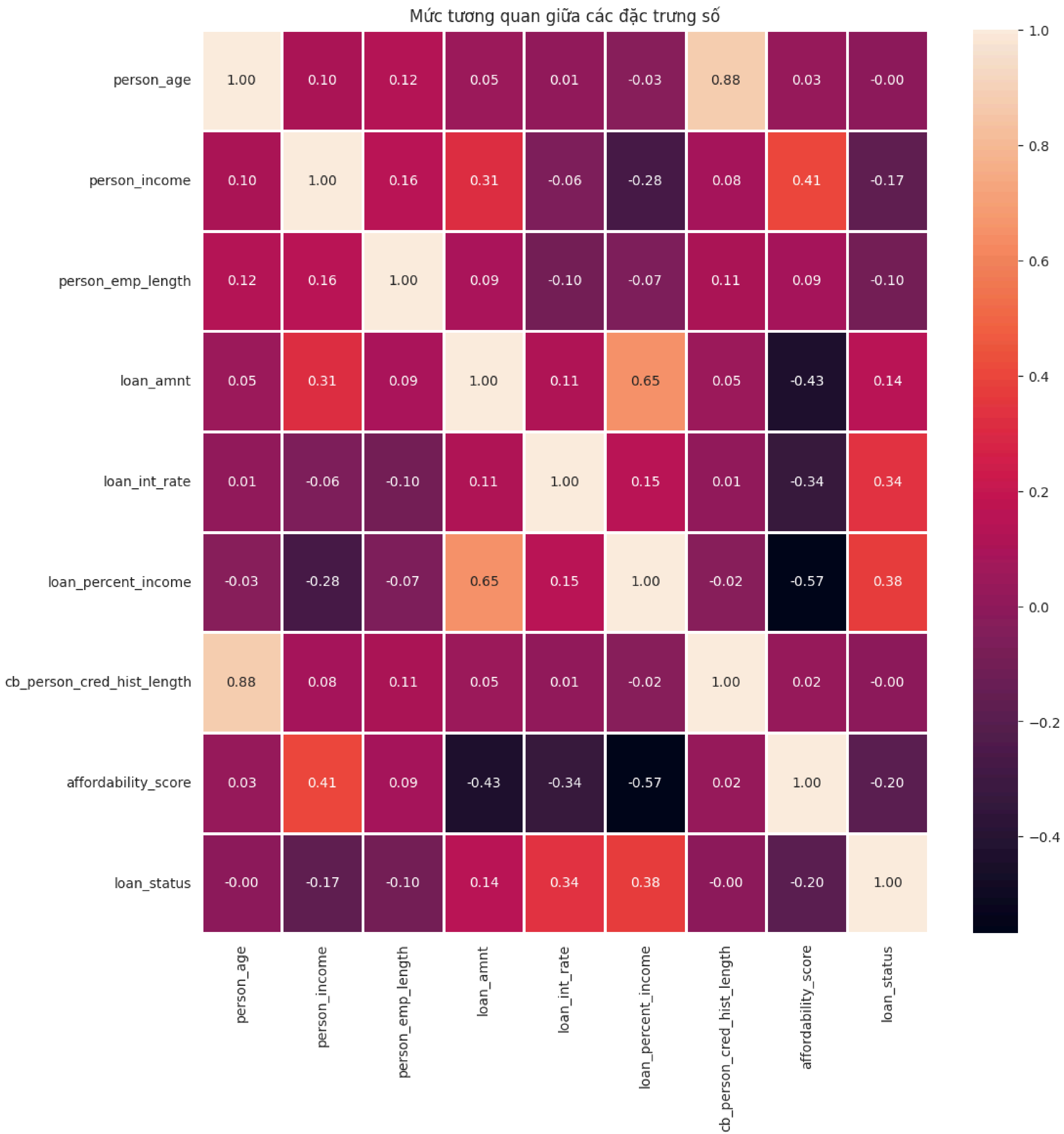
XGBOOST CLASSIFICATION FEATURE IMPORTANCE

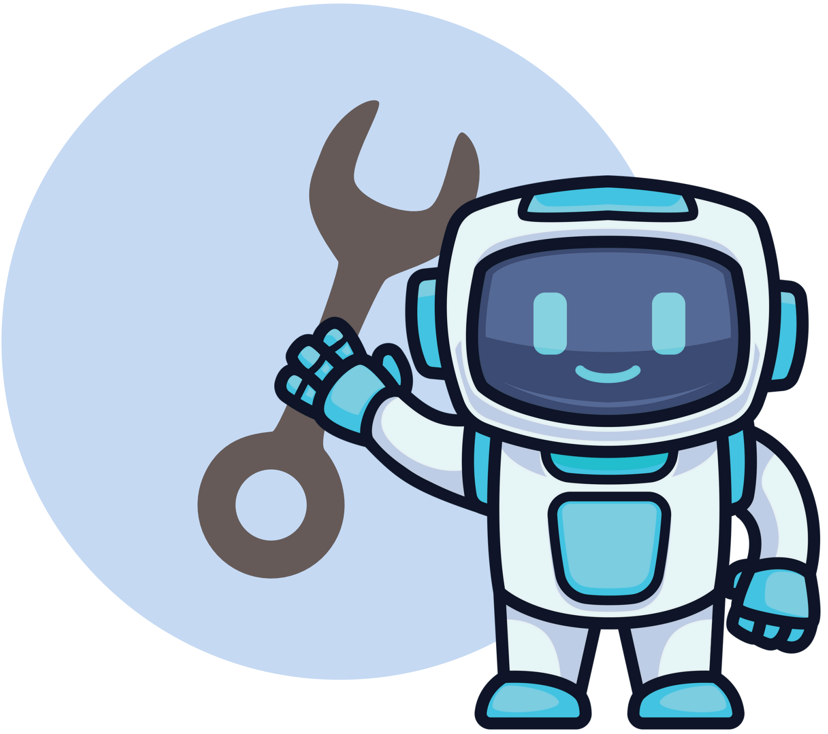


Analysis of Variance (ANNOVA)

Nhóm	Đặc trưng (Features)
Nhóm 1 (Quan trọng cao)	loan_int_rate affordability_score (đặc trưng tự tạo) loan_percent_income home_ownership_RENT available_funds_ratio (đặc trưng tự tạo)
Nhóm 2 (Quan trọng trung bình)	person_income loan_grade_encoded person_emp_length loan_amnt loan_intent_HOMEIMPROVEMENT loan_intent_MEDICAL loan_intent_PERSONAL
Nhóm 3 (Quan trọng thấp)	person_age loan_intent_EDUCATION cb_person_default_on_file_encoded loan_intent_VENTURE home_ownership_OWN home_ownership_OTHER cb_person_cred_hist_length

FEATURE ENGINEERING





BUILD & TUNING MODEL

LOGISTICS REGRESSION

F1 - score	Accuracy
0.7233	0.8173

RANDOM FOREST

F1 - score	Accuracy
0.8877	0.95

=> THE MODELS THAT FIT NONLINEAR DATA WILL BE MORE SUITABLE FOR THIS DATA

BASELINE MODEL SELECTION

CRITERIA FOR MODEL SELECTION

Suitable for:

- Large datasets
- Nonlinear data
- imbalanced datasets

BASELINE MODEL SELECTION

OPTIMIZE PARAMETER

Find the best hyperparameters

LIGHTGBM

CATBOOST

XGBOOST

K-FOLD VALIDATION

CV = 5

+

GRIDSEARCHCV

Identifying the best hyperparameter range

↓

OPTUNA

Deep fine-tuning

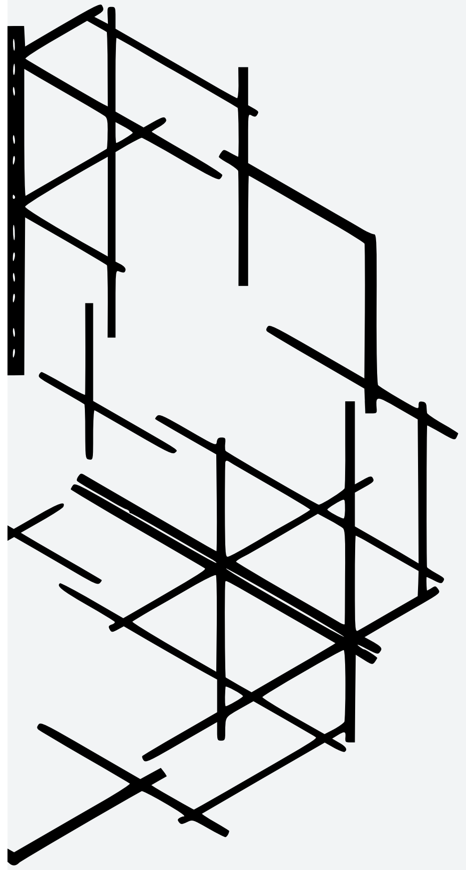
FIND THE BEST HYPERPARAMETERS

Performance results table of the model

Models	F1-score	Accuracy
Logistic Regression	0.7233	0.8173
Random Forest	0.8877	0.95
LightGBM	0.8832	0.945
Catboost	0.888	0.9482
Xgboost	0.8928	0.9518

RESULTS

ENSEMBLE MODELS



IMPROVING ACCURACY

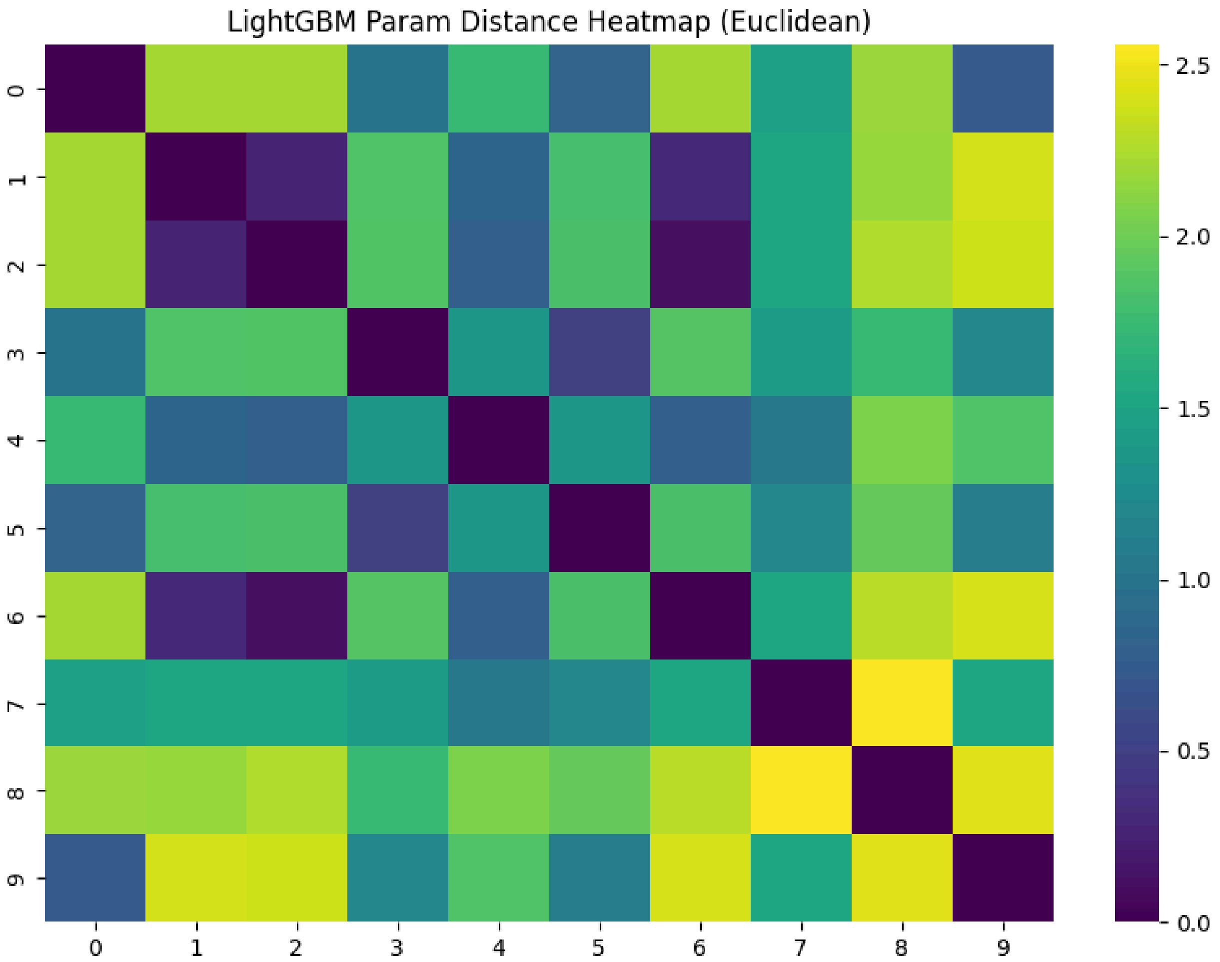
IMPROVING STABILITY

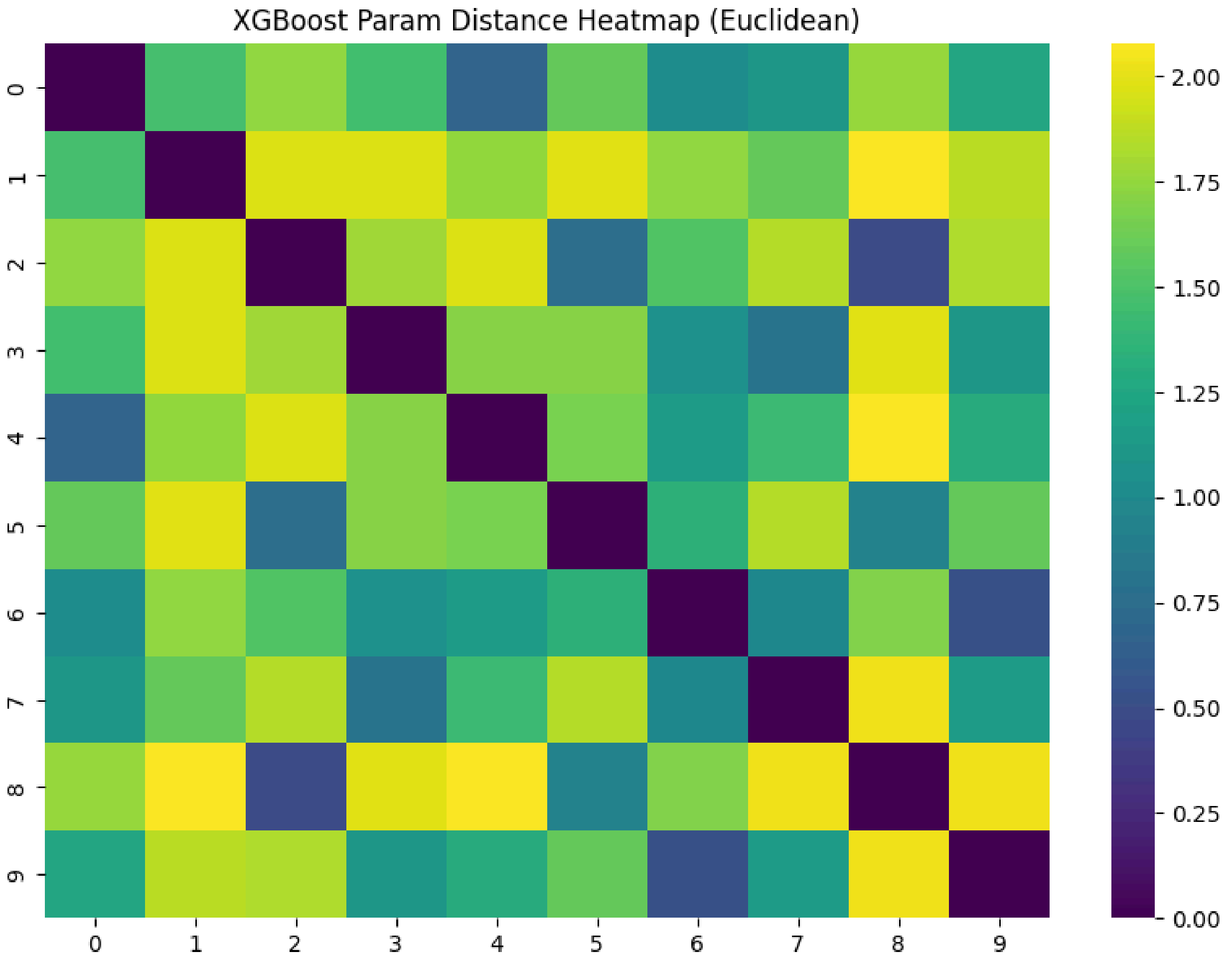
IMPROVING GENERALIZATION

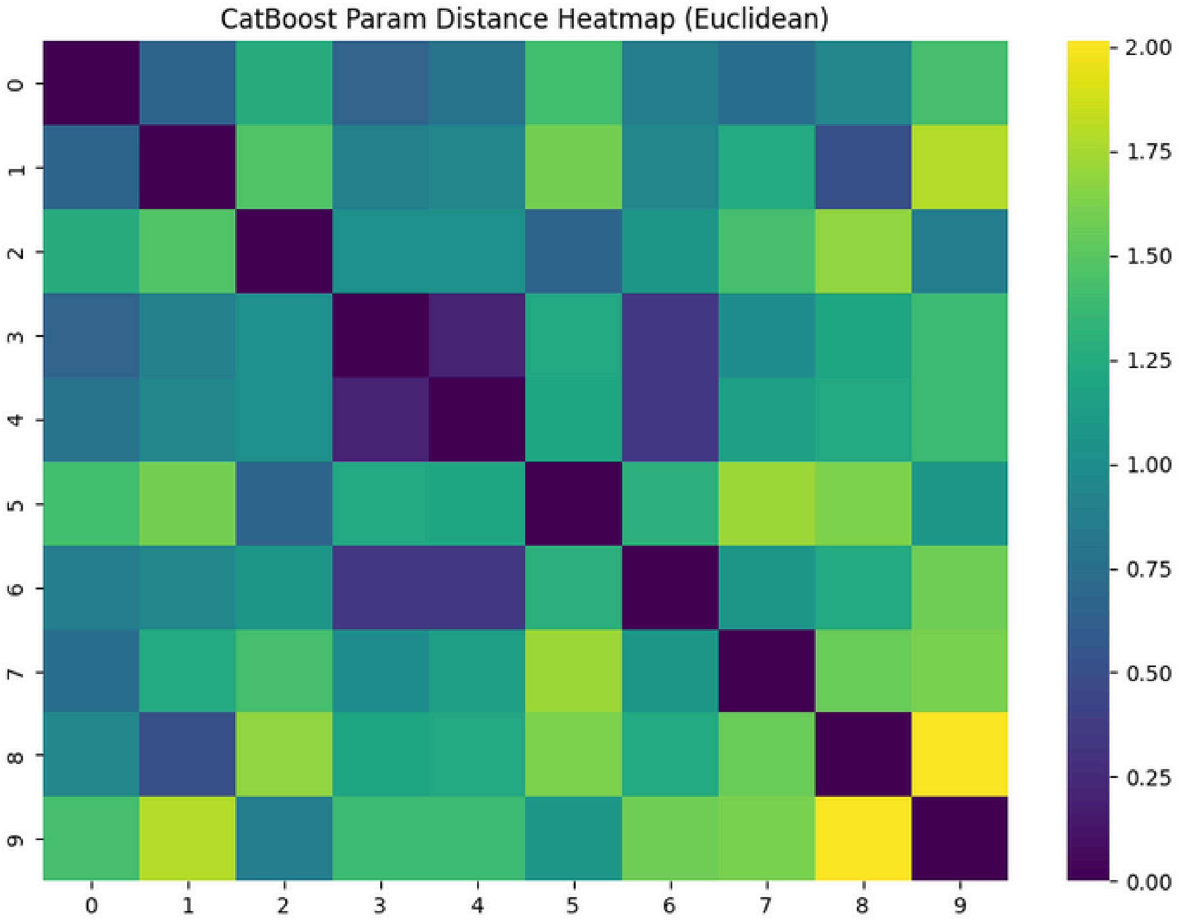
ENSEMBLE MODEL

Models	F1-score	Accuracy	Chọn
Logistic Regression	0.7233	0.8173	X
Random Forest	0.8877	0.95	X
LightGBM	0.8832	0.945	10 best sets of hyperparameters
Catboost	0.888	0.9482	10 best sets of hyperparameters
Xgboost	0.8928	0.9518	10 best sets of hyperparameters

RESOURCE PREPARATION







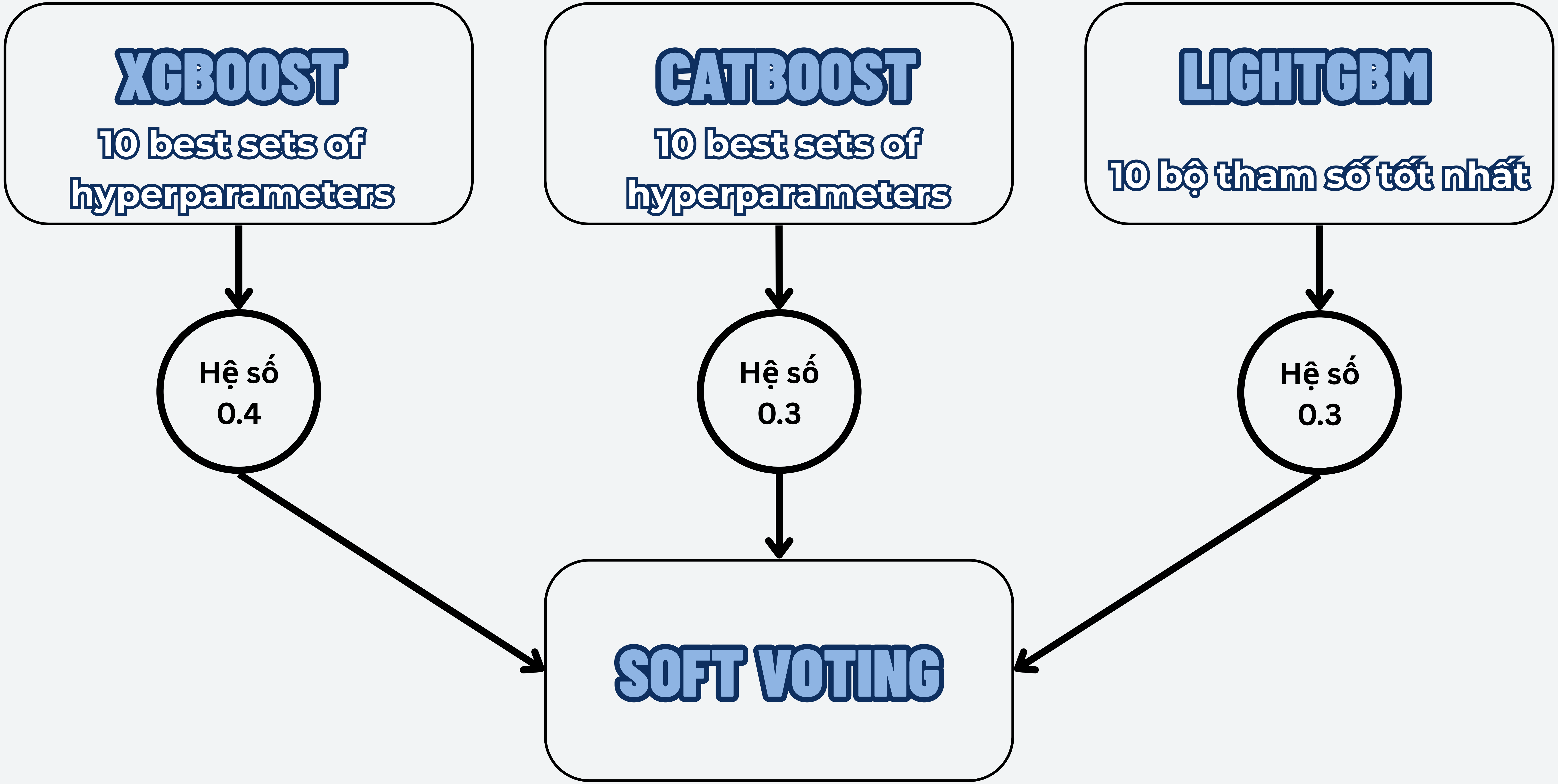
Euclidean distance of LightGBM: 1.5660 [0, 2.6458]

Euclidean distance of Xgboost: 1.4961 [0, 3.3166]

Euclidean distance of Catboost: 1.1407 [0, 2.6458]

=> THE CURRENT HYPERPARAMETER SETS ARE ACCEPTABLE

CHECKING HYPERPARAMETER SETS


XGBOOST
10 best sets of hyperparameters
CATBOOST
10 best sets of hyperparameters
LIGHTGBM
10 bộ tham số tốt nhất
Hệ số 0.4
Hệ số 0.3
Hệ số 0.3
SOFT VOTING

SOFT VOTING

Models	F1-score	Accuracy
Logistic Regression	400	reg_lambda
Random Forest	0.8877	0.95
LightGBM	0.8832	0.945
Catboost	0.888	0.9482
Xgboost	0.8928	0.9518
Soft Voting(Xgboost, Catboost, LightGBM)(best hyperparameter)	0.8934	0.9521
Stacking(Xgboost, Catboost, LightGBM)(best hyperparameter)	0.872	0.9354
Soft voting(Xgboost*10, Catboost*10, LightGBM*10)	0.894	0.9524

KẾT QUẢ

THANK YOU!