

BÁO CÁO PHÂN TÍCH DỮ LIỆU

PHÂN TÍCH VÀ DỰ ĐOÁN RỦI RO TÍN DỤNG

Giáo viên hướng dẫn : Hồ Hướng Thiên

Thành viên nhóm:

1. Lương Minh Thông - 2254050064
2. Nguyễn Thị Hạnh Quyên - 2254050056
3. Nguyễn Thị Phương Oanh - 2254052057

Nội dung

- 01 Giới thiệu
- 02 Tiền xử lí dữ liệu
- 03 Trực quan hóa dữ liệu
- 04 Ma trận tương quan
- 05 Mô hình thuật toán & đánh giá
- 06 Xây dựng mô hình dự đoán rủi ro tín dụng trên Streamlit
- 07 Kết luận

Giới thiệu

Mục tiêu nghiên cứu

- Tìm hiểu các yếu tố ảnh hưởng xấu đến rủi ro tín dụng.
- Phân tích xu hướng và hành vi tín dụng của người vay dựa trên bộ dữ liệu nghiên cứu để xác định các đặc điểm chính liên quan đến rủi ro tín dụng từ đó hỗ trợ việc ra quyết định phê duyệt khoản vay.
- Đánh giá hiệu suất của các thuật toán học máy (machine learning) áp dụng cho dự đoán rủi ro tín dụng.
- Xây dựng mô hình dự đoán người vay có khả năng vỡ nợ hay không dựa trên việc sử dụng phương pháp học máy machine learning.

Giới thiệu

Bộ dữ liệu

	person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
0	22	59000	RENT	123.0	PERSONAL	D	35000	16.02	1	0.59	Y	3
1	21	9600	OWN	5.0	EDUCATION	B	1000	11.14	0	0.10	N	2
2	25	9600	MORTGAGE	1.0	MEDICAL	C	5500	12.87	1	0.57	N	3
3	23	65500	RENT	4.0	MEDICAL	C	35000	15.23	1	0.53	N	2
4	24	54400	RENT	8.0	MEDICAL	C	35000	14.27	1	0.55	Y	4
...
32576	57	53000	MORTGAGE	1.0	PERSONAL	C	5800	13.16	0	0.11	N	30
32577	54	120000	MORTGAGE	4.0	PERSONAL	A	17625	7.49	0	0.15	N	19
32578	65	76000	RENT	3.0	HOMEIMPROVEMENT	B	35000	10.99	1	0.46	N	28
32579	56	150000	MORTGAGE	5.0	PERSONAL	B	15000	11.48	0	0.10	N	26
32580	66	42000	RENT	2.0	MEDICAL	B	6475	9.99	0	0.15	N	30

32581 rows × 12 columns

Giới thiệu

Mô tả dữ liệu

Cột	Mô tả
person_age	Tuổi của cá nhân nộp đơn xin vay.
person_income	Thu nhập hàng năm của cá nhân.
person_home_ownership	Tình trạng sở hữu nhà của người vay. RENT: Thuê nhà OWN: Sở hữu nhà MORTGAGE: Có thế chấp nhà
person_emp_length	Số năm người nộp đơn đang làm việc.

loan_intent	Mục đích vay vốn của người nộp đơn: EDUCATION: Giáo dục MEDICAL: Y tế PERSONAL: Cá nhân VENTURE: Khởi nghiệp HOMEIMPROVEMENT: Cải thiện nhà cửa DEBTCONSOLIDATION: Hợp nhất nợ (Thanh toán cho nhiều khoản nợ khác)
-------------	---

Giới thiệu

Mô tả dữ liệu

loan_grade	<p>Điểm tín dụng được chỉ định cho khoản vay dựa trên mức độ tín nhiệm (xét với các yếu tố: lịch sử tín dụng, chất lượng tài sản thế chấp, khả năng trả nợ,...) của người vay:</p> <p>A: Người vay có độ tín nhiệm cao, cho thấy rủi ro thấp.</p> <p>B: Người vay có rủi ro tương đối thấp, nhưng không có độ tín nhiệm cao như mức A.</p> <p>C: Độ tín nhiệm của người vay ở mức trung bình.</p> <p>D: Người vay được coi là có rủi ro cao hơn so với các mức trước đó.</p> <p>E: Độ tín nhiệm của người vay thấp hơn, cho thấy rủi ro cao hơn.</p> <p>F: Người vay có rủi ro tín dụng đáng kể.</p> <p>G: Độ tín nhiệm của người vay là thấp nhất, biểu thị rủi ro cao nhất.</p>
------------	---

loan_amnt	Số tiền vay.
loan_int_rate	Lãi suất áp dụng cho khoản vay.
loan_percent_income	Tỷ lệ phần trăm số tiền vay theo tổng thu nhập.
cb_person_default_on_file	Lịch sử vỡ nợ của cá nhân theo hồ sơ của cơ quan tín dụng: Y: Cá nhân có lịch sử nợ xấu trong hồ sơ tín dụng. N: Cá nhân này không có tiền sử vi phạm nợ xấu.
cb_person_cred_hist_length	Số năm lịch sử cá nhân kể từ khoản vay đầu tiên.
loan_status	Trạng thái của khoản vay (biến mục tiêu): 0: Không vỡ nợ - Người vay trả nợ thành công theo đúng thỏa thuận và không xảy ra vỡ nợ. 1: Vỡ nợ - Người vay không trả nợ đúng hạn theo các điều khoản đã thỏa thuận và vỡ nợ khoản vay.

Tiền xử lý dữ liệu - Làm sạch dữ liệu

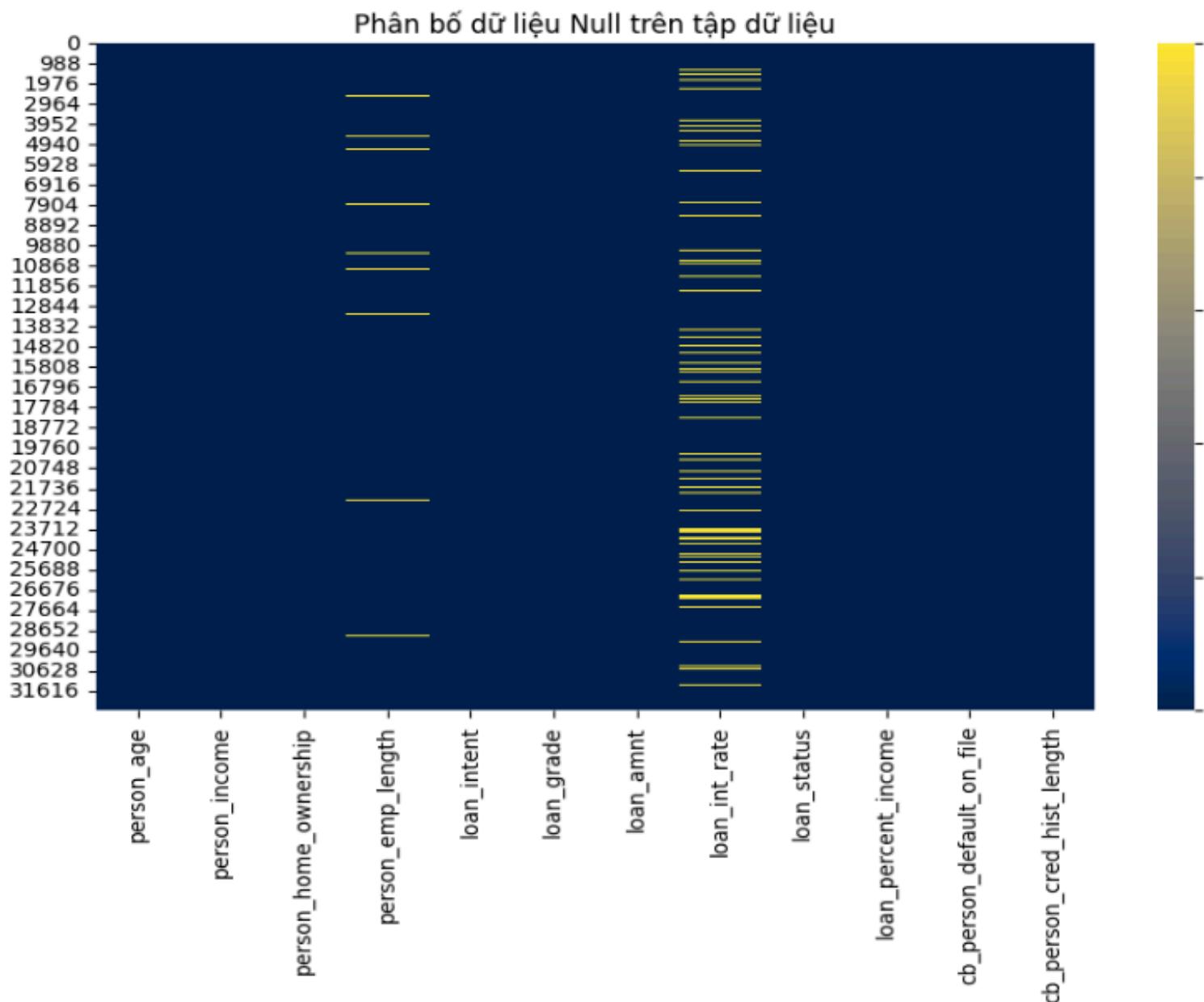


Giá trị rỗng (Null Values)

Kiểm tra giá trị Null:

person_age	0
person_income	0
person_home_ownership	0
person_emp_length	895
loan_intent	0
loan_grade	0
loan_amnt	0
loan_int_rate	3116
loan_status	0
loan_percent_income	0
cb_person_default_on_file	0
cb_person_cred_hist_length	0

Phân bố của giá trị Null:



Sau khi loại bỏ giá trị Null:

person_age	0
person_income	0
person_home_ownership	0
person_emp_length	0
loan_intent	0
loan_grade	0
loan_amnt	0
loan_int_rate	0
loan_status	0
loan_percent_income	0
cb_person_default_on_file	0
cb_person_cred_hist_length	0

Tiền xử lý dữ liệu - Làm sạch dữ liệu

• • • •

Giá trị trùng lặp (Duplicate Rows)

Kiểm tra giá trị trùng lặp:

```
df.duplicated().sum()
```

137

Sau khi loại bỏ hàng trùng lặp:

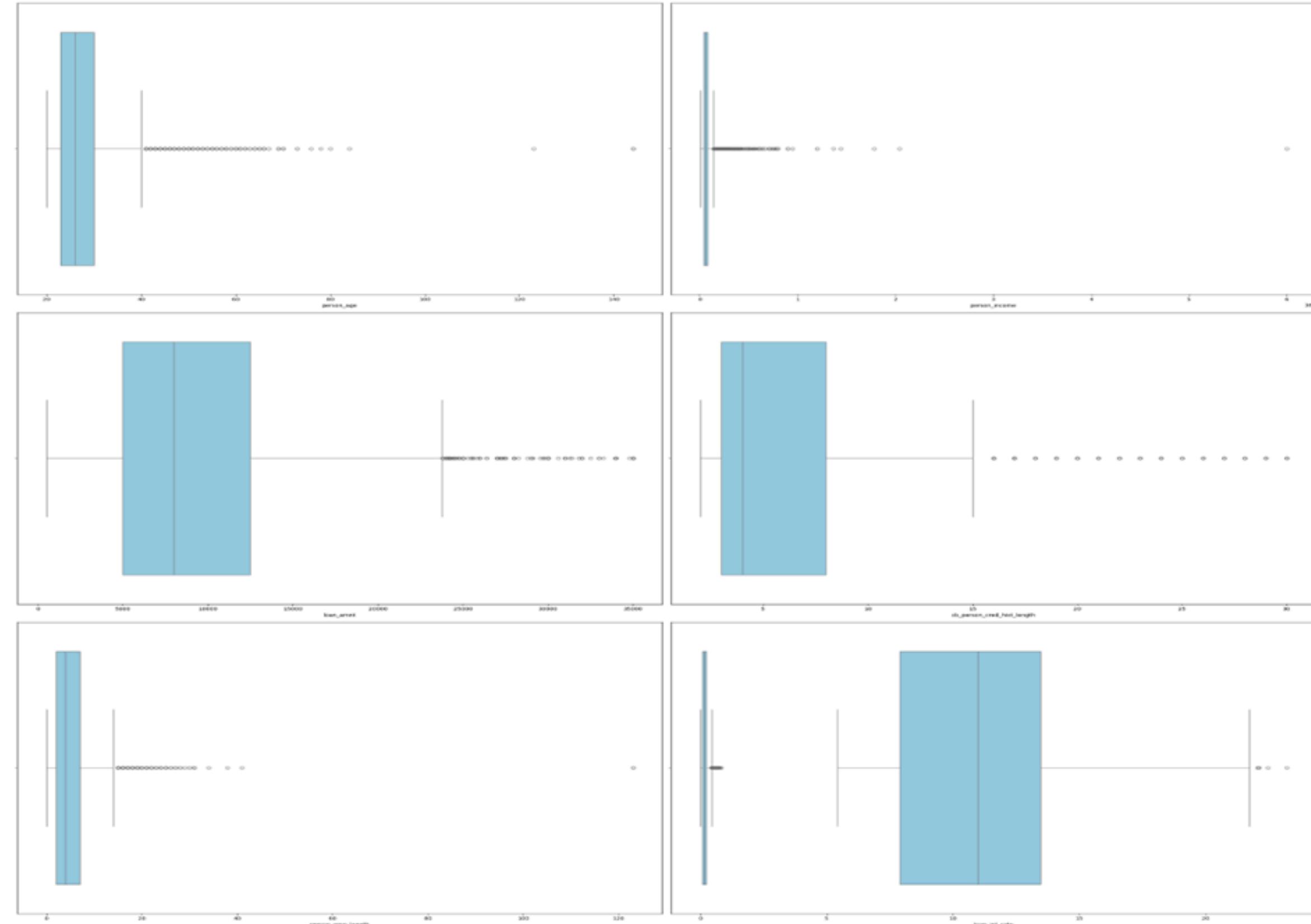
```
df.shape
```

(28501, 12)

Tiền xử lý dữ liệu - Làm sạch dữ liệu



Giá trị ngoại lệ (Outlier Values)

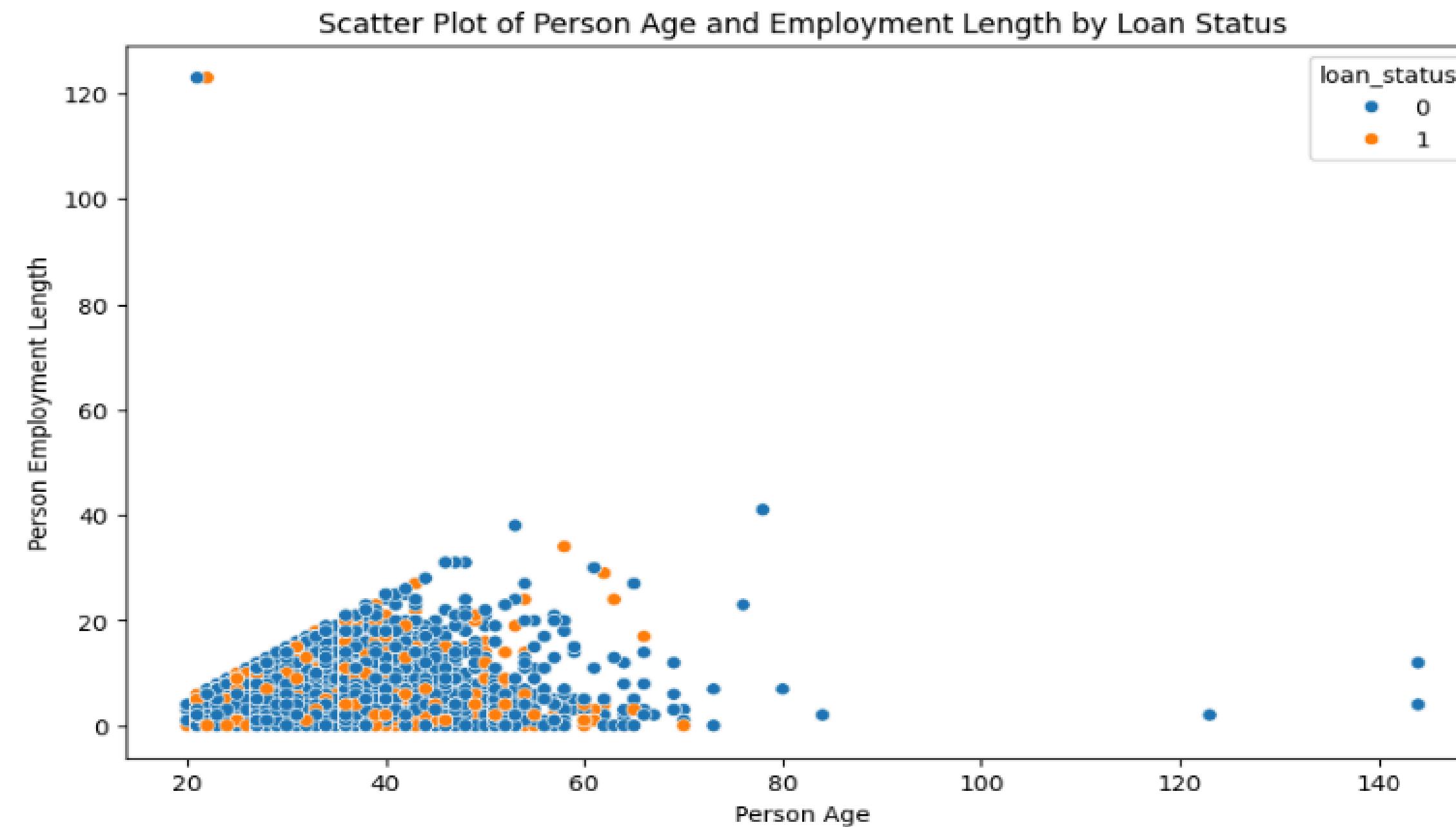


Tiền xử lý dữ liệu - Làm sạch dữ liệu



Giá trị ngoại lệ (Outlier Values)

Hai đặc trưng person_age và person_emp_length có phân bố ở các khoảng giá trị chưa hợp lý cần loại bỏ:

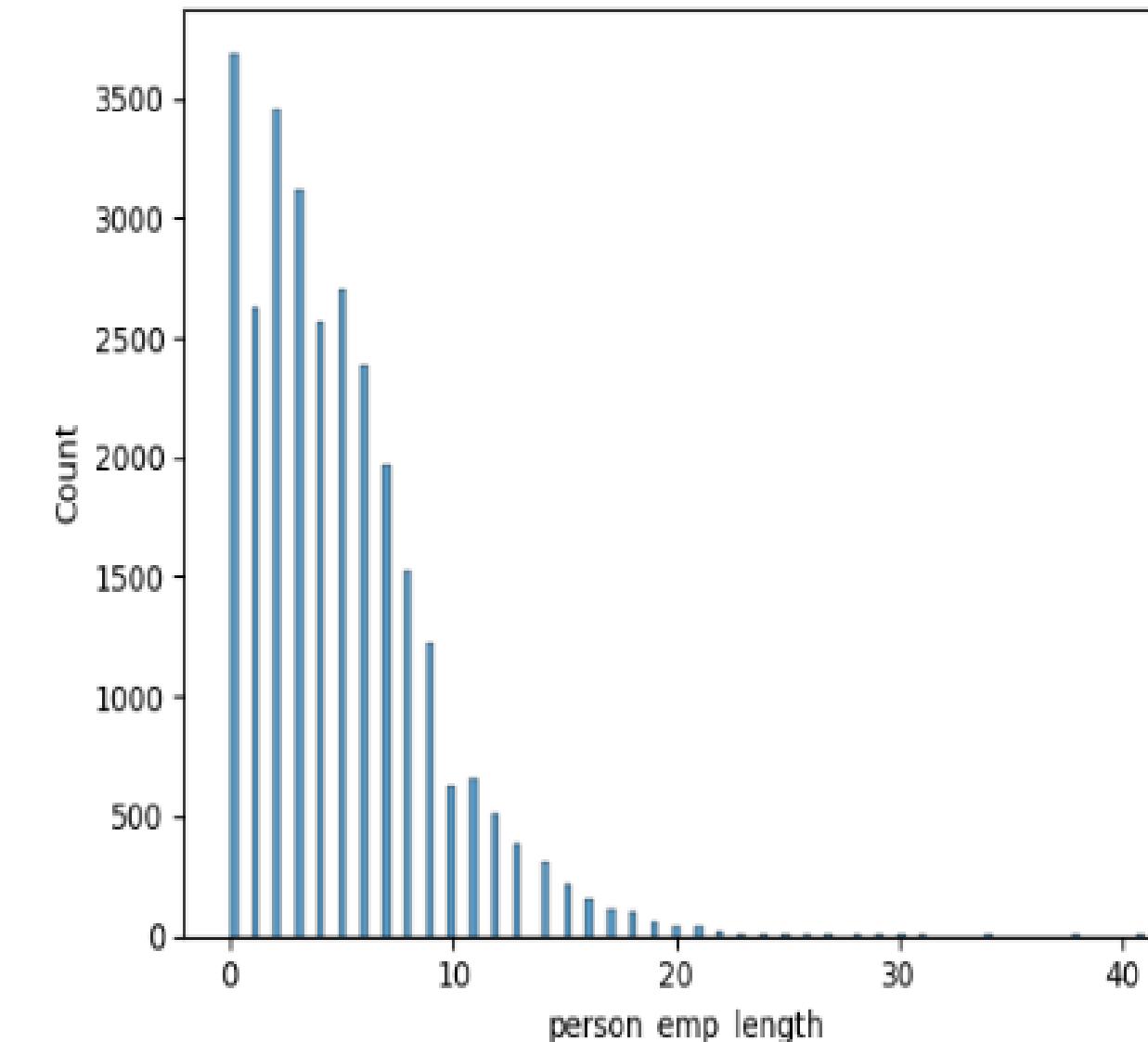
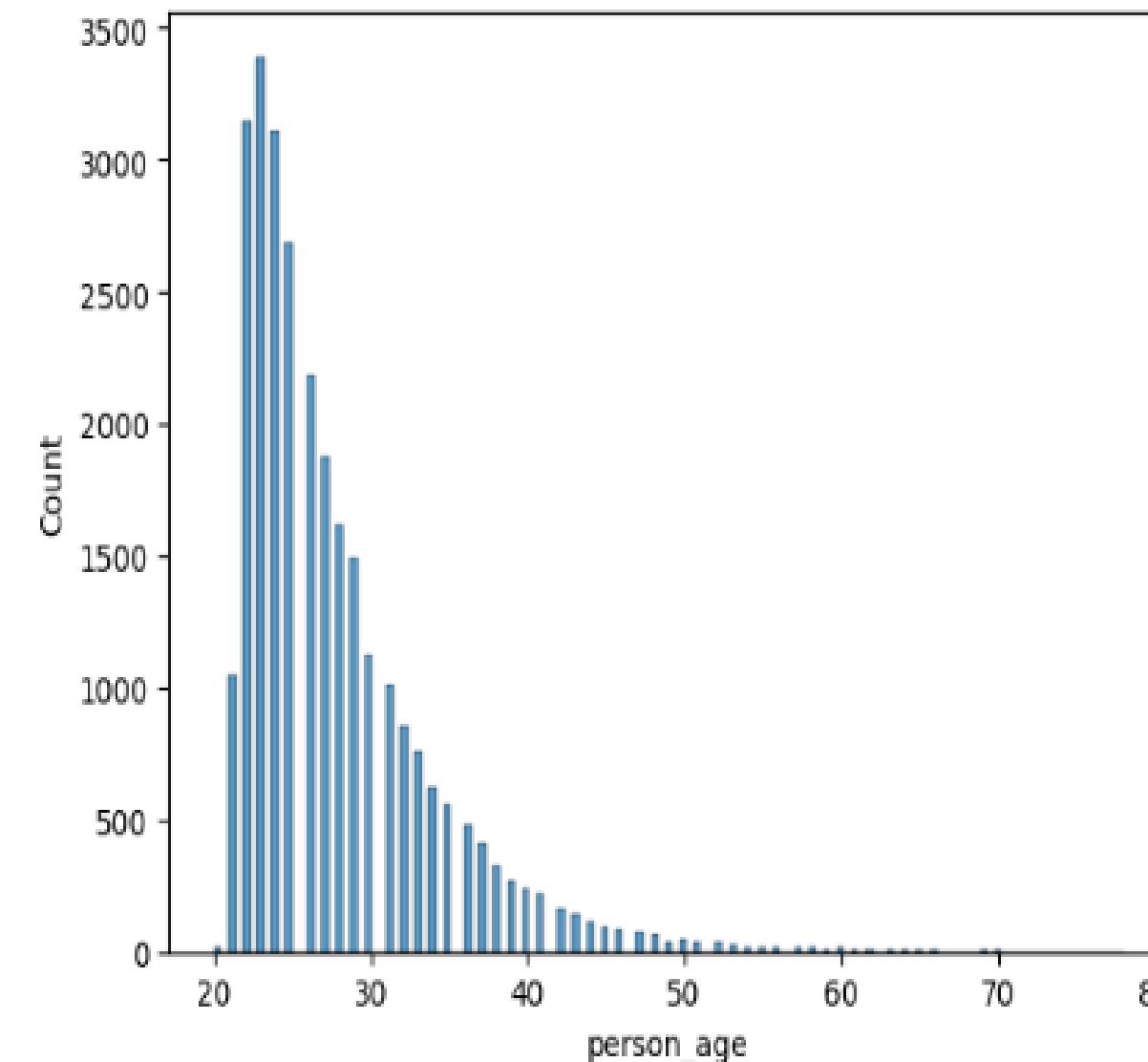


Tiền xử lý dữ liệu - Làm sạch dữ liệu



Giá trị ngoại lệ (Outlier Values)

- Sau khi loại bỏ giá trị Outliers:



Tiền xử lý dữ liệu - Chuyển đổi dữ liệu



Chuyển các dữ liệu cột sang dạng mã hóa

```
# OneHot encoding categorical variables
num_col = df.select_dtypes(exclude = 'object')
char_col = df.select_dtypes(include = 'object')

encoded_char_col = pd.get_dummies(char_col)

df = pd.concat([num_col, encoded_char_col], axis=1)
df.head()
```

- Sau khi chuyển đổi

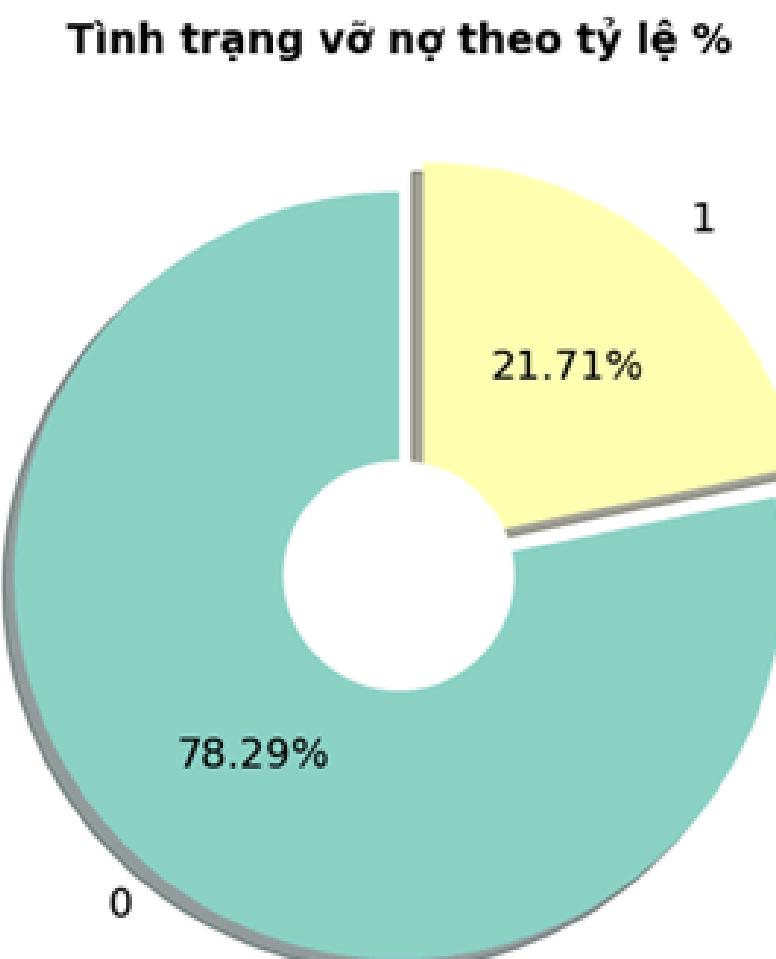
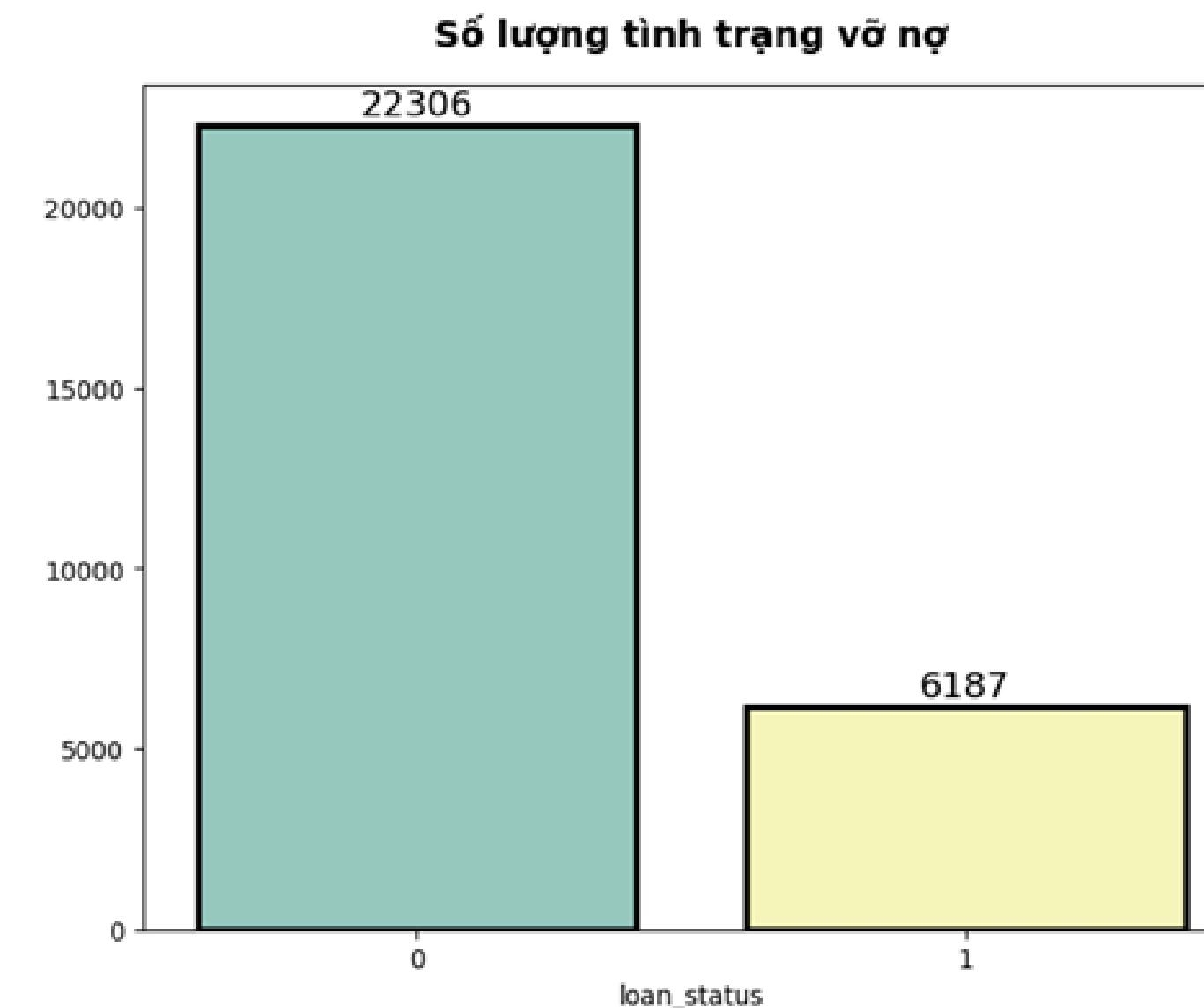
	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_cred_hist_length	person_home_ownership_MORTGAGE	person_home_ownership_OTHER	...	loan_intent_VENTURE	
1	21	9600	5.0	1000	11.14	0	0.10	2	False	False	...	False	
2	25	9600	1.0	5500	12.87	1	0.57	3	True	False	...	False	
3	23	65500	4.0	35000	15.23	1	0.53	2	False	False	...	False	
4	24	54400	8.0	35000	14.27	1	0.55	4	False	False	...	False	
5	21	9900	2.0	2500	7.14	1	0.25	2	False	False	...	True	

5 rows × 27 columns

Tiền xử lý dữ liệu - Chuyển đổi dữ liệu



Xử lý dữ liệu mất cân bằng



Dữ liệu có sự mất cân bằng tỷ lệ gần 4:1

Tiền xử lý dữ liệu - Chuyển đổi dữ liệu

Xử lý dữ liệu mất cân bằng



Phương pháp Oversampling tập trung vào giải quyết hiện tượng mất cân bằng mẫu bằng cách gia tăng kích thước mẫu thuộc lớp thiểu số.

- Hai phương pháp chính để thực hiện Oversampling:
 - + **Lập lại mẫu hiện có để tăng số lượng mẫu**
 - + **Tạo mẫu mới dựa trên tổng hợp của các mẫu cũ (SMOTE)**: Thuật toán chọn 2 hay nhiều trường hợp giống nhau (sử dụng thước đo khoảng cách để so sánh) và xáo trộn một cá thể một thuộc tính tại một thời điểm bằng một lượng ngẫu nhiên trong khoảng chênh lệch với các trường hợp lân cận.

Tiền xử lý dữ liệu - Chuyển đổi dữ liệu

Xử lý dữ liệu mất cân bằng - Kỹ thuật SMOTE

- Sau khi xử lý dữ liệu đã cân bằng

```
from imblearn.over_sampling import SMOTE
smote = SMOTE(sampling_strategy='minority')
x, y = smote.fit_resample(x,y)
y.value_counts()
```

```
loan_status
0    22306
1    22306
Name: count, dtype: int64
```

Tiền xử lý dữ liệu - Chuyển đổi dữ liệu

Đào tạo và phân chia dữ liệu cho machine learning

```
x = df.drop('loan_status', axis=1)  
y = df['loan_status']
```

```
from sklearn.model_selection import train_test_split  
# split train and test sets  
X_train, X_test, y_train, y_test = train_test_split(  
                                    df.drop(labels=['loan_status'], axis=1),  
                                    df['loan_status'],  
                                    test_size=0.3,  
                                    random_state=0)
```

```
X_train.shape, X_test.shape
```

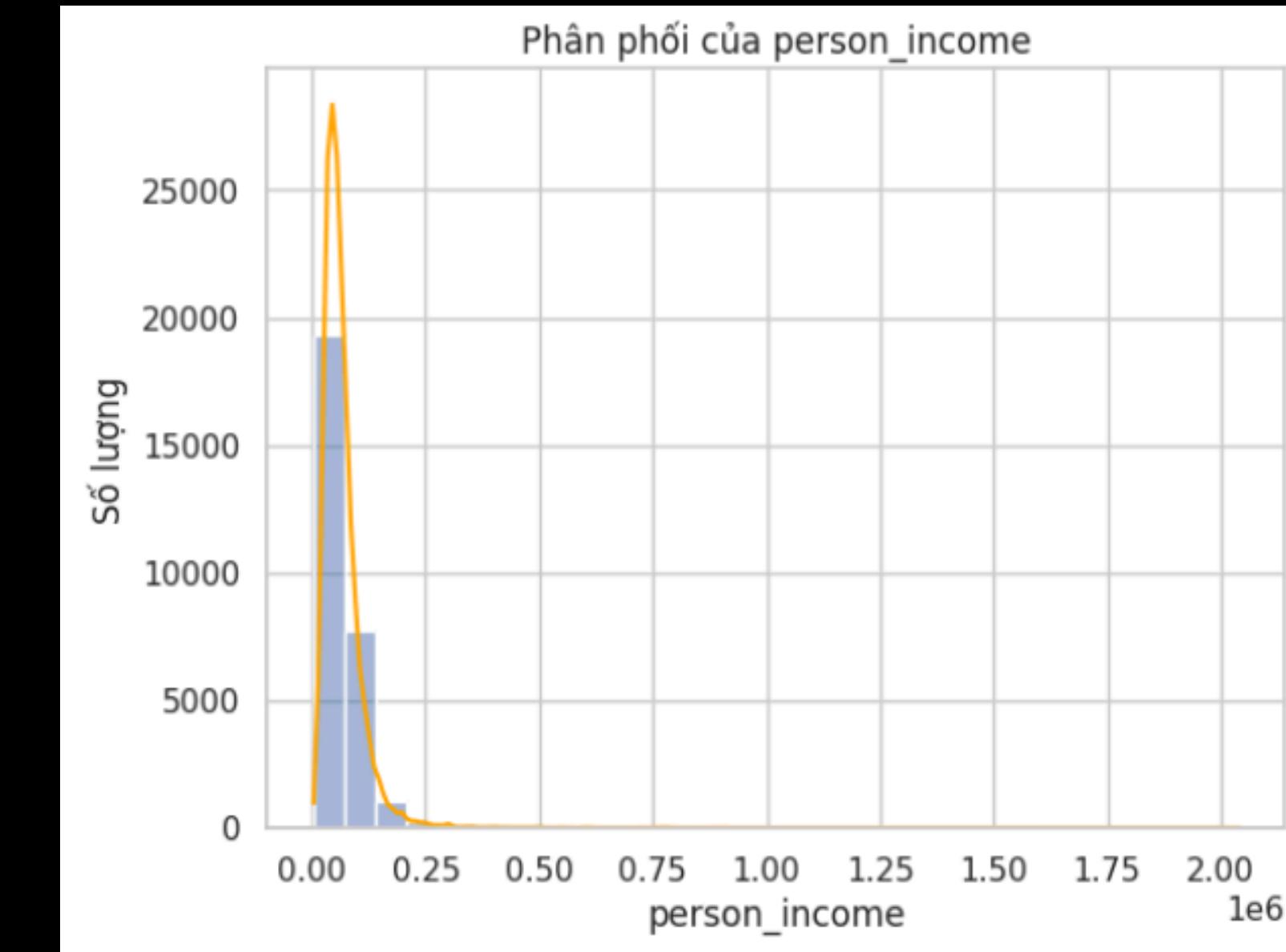
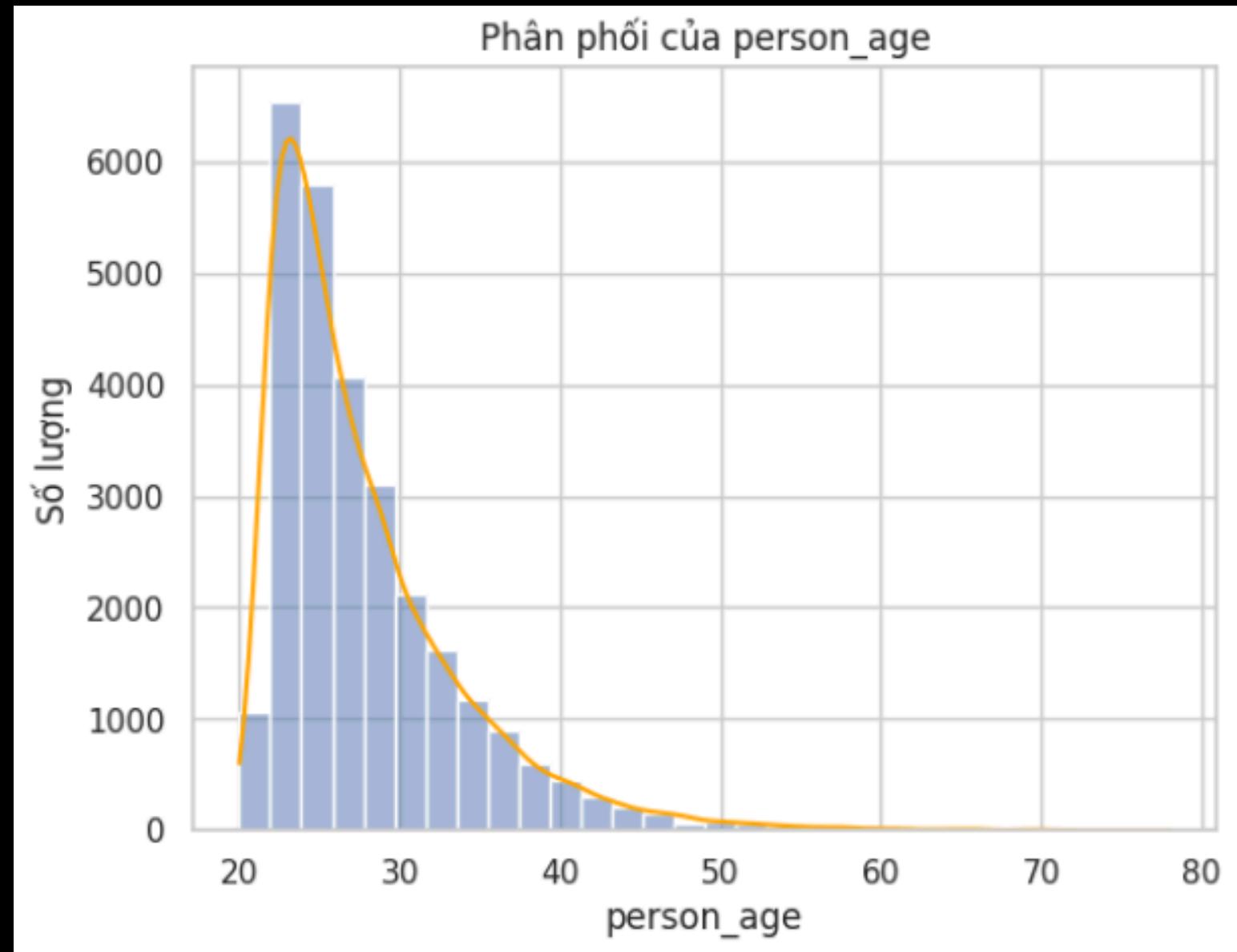
```
((19945, 26), (8548, 26))
```

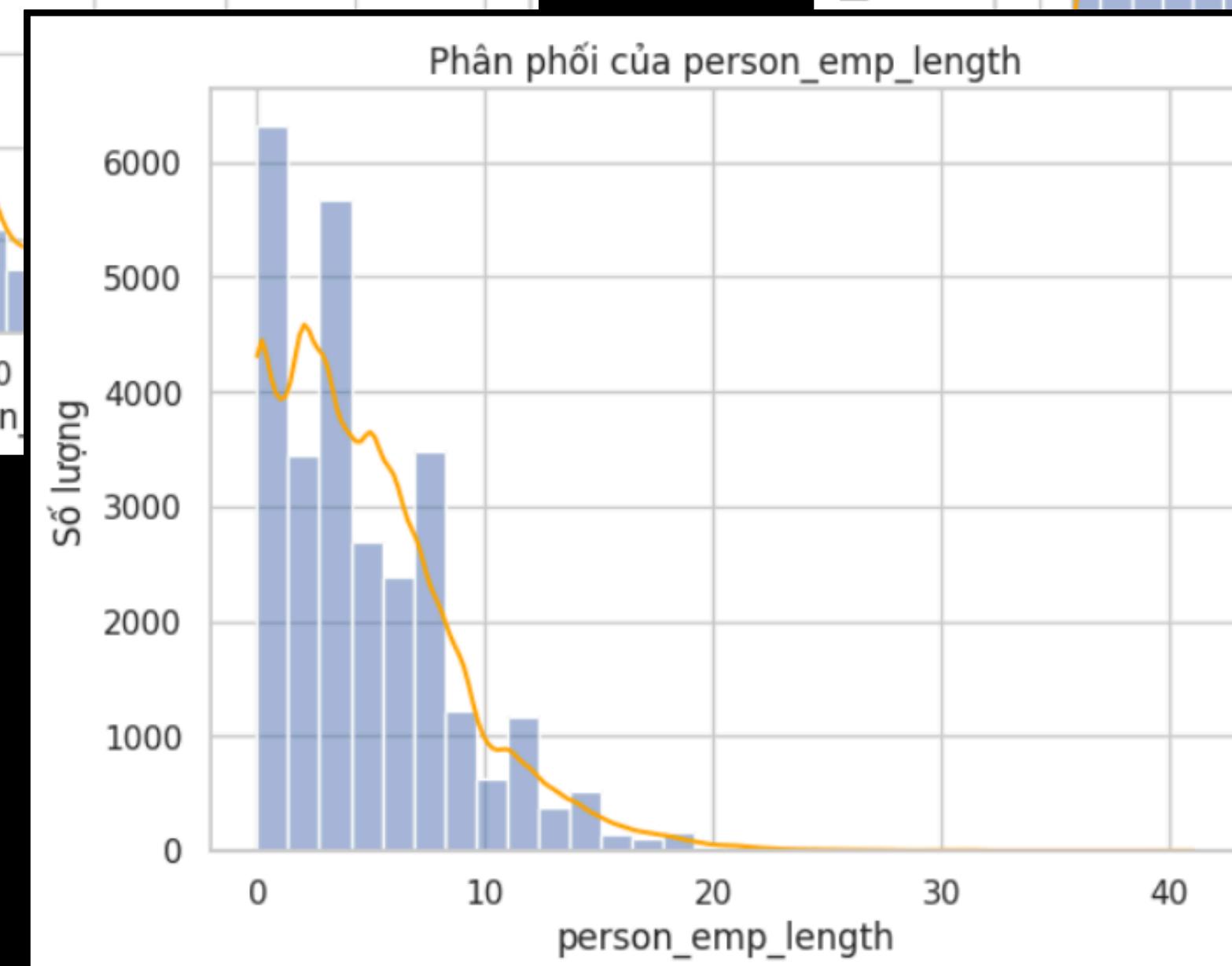
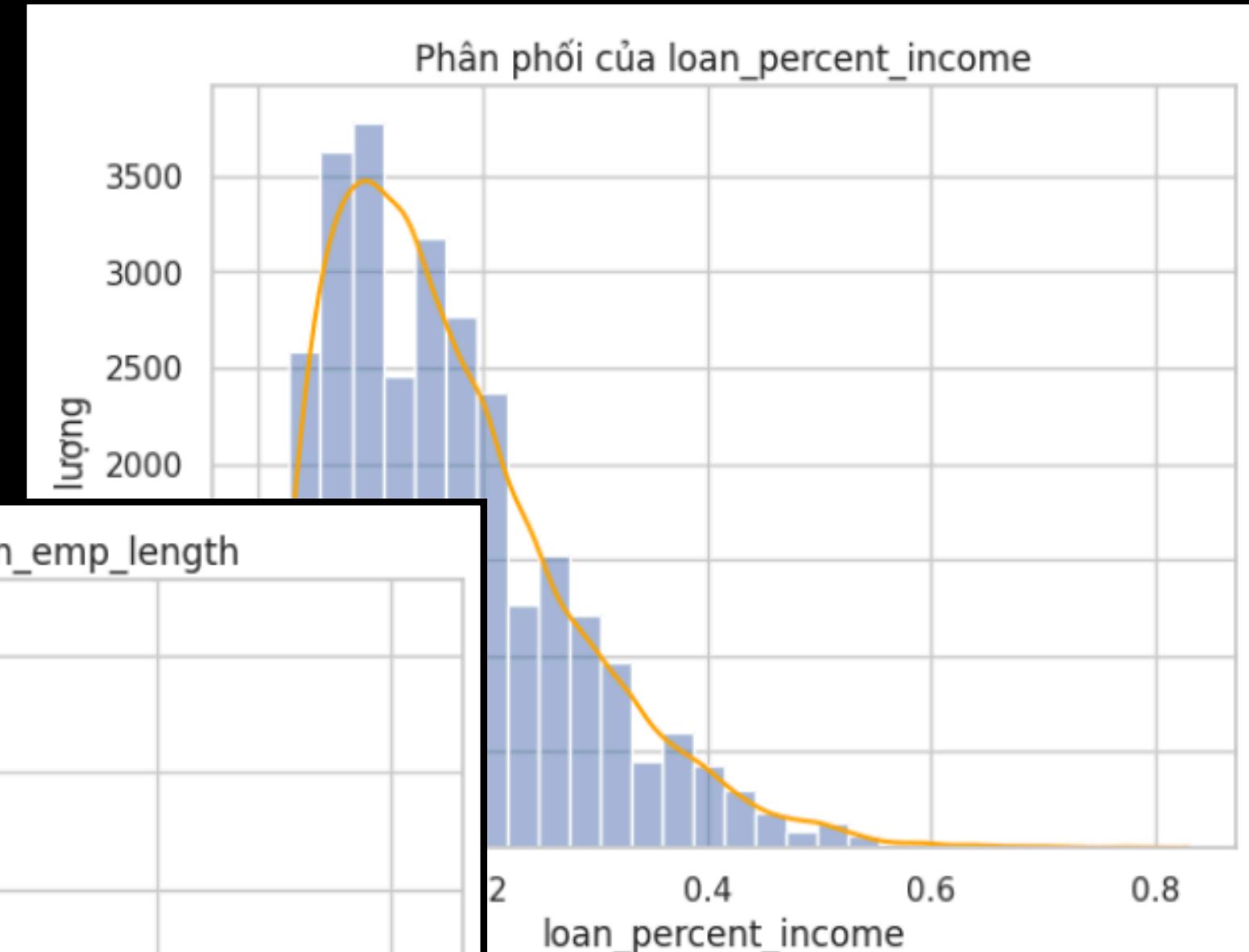
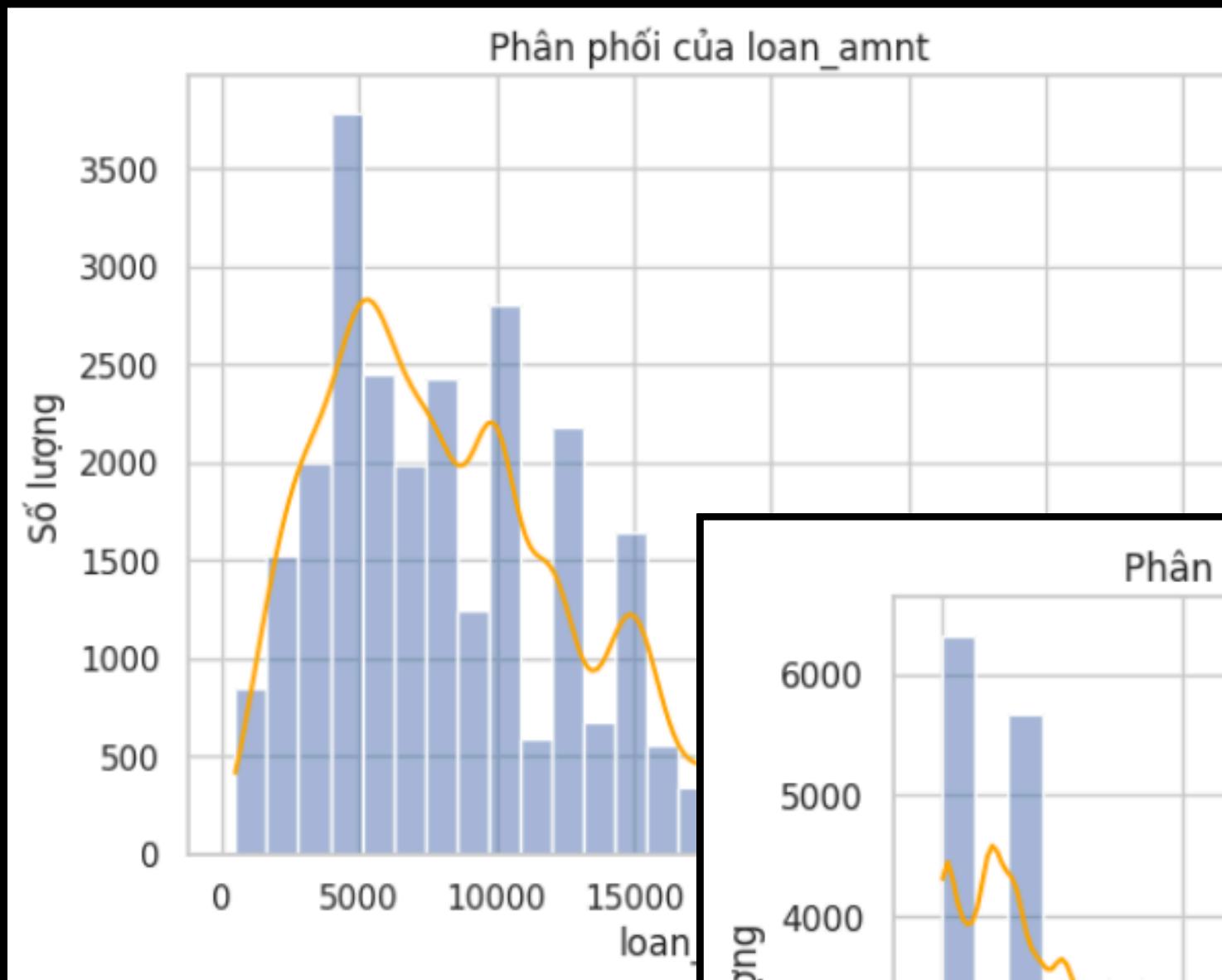
```
y_train.shape, y_test.shape
```

```
((19945,), (8548,))
```

Trực quan hóa dữ liệu

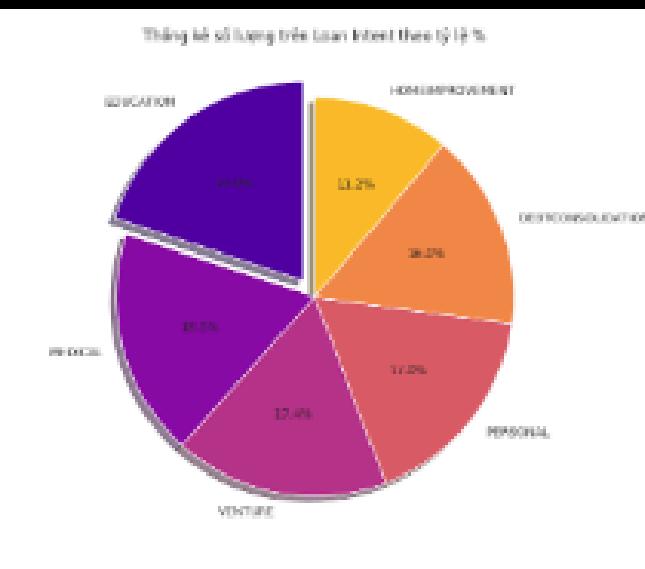
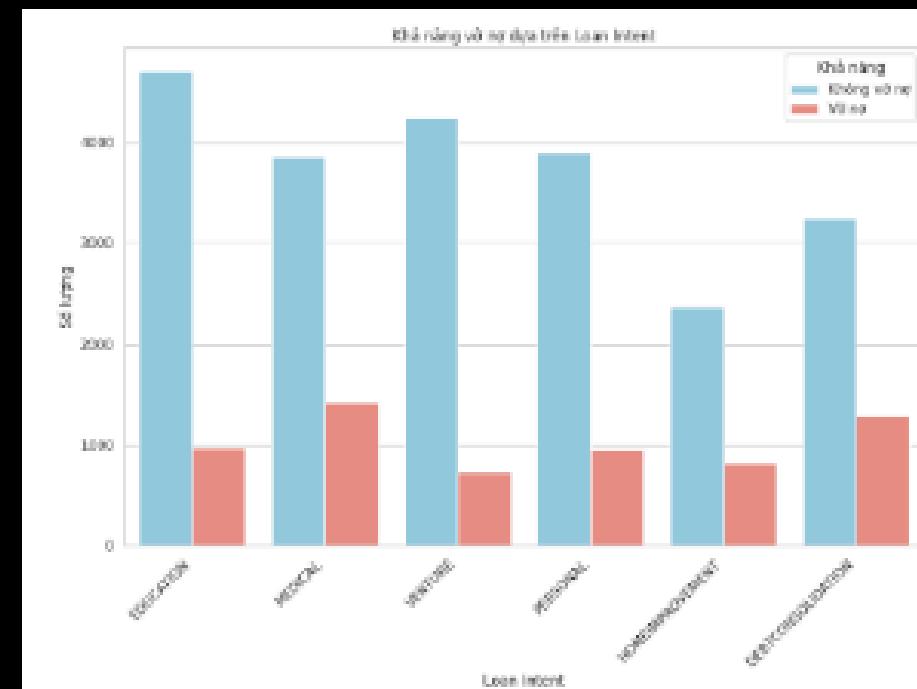
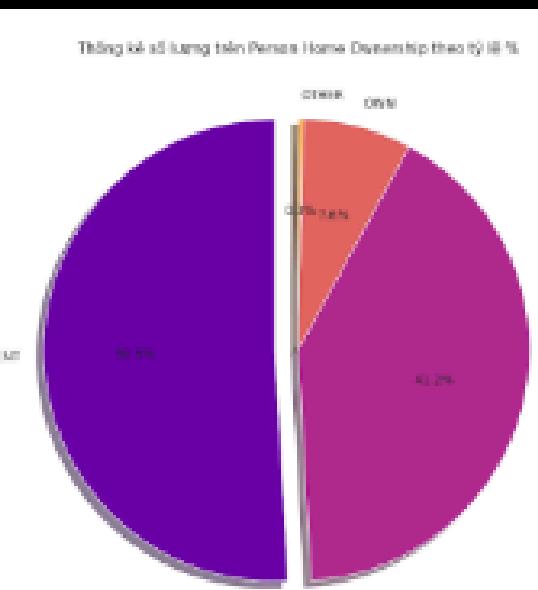
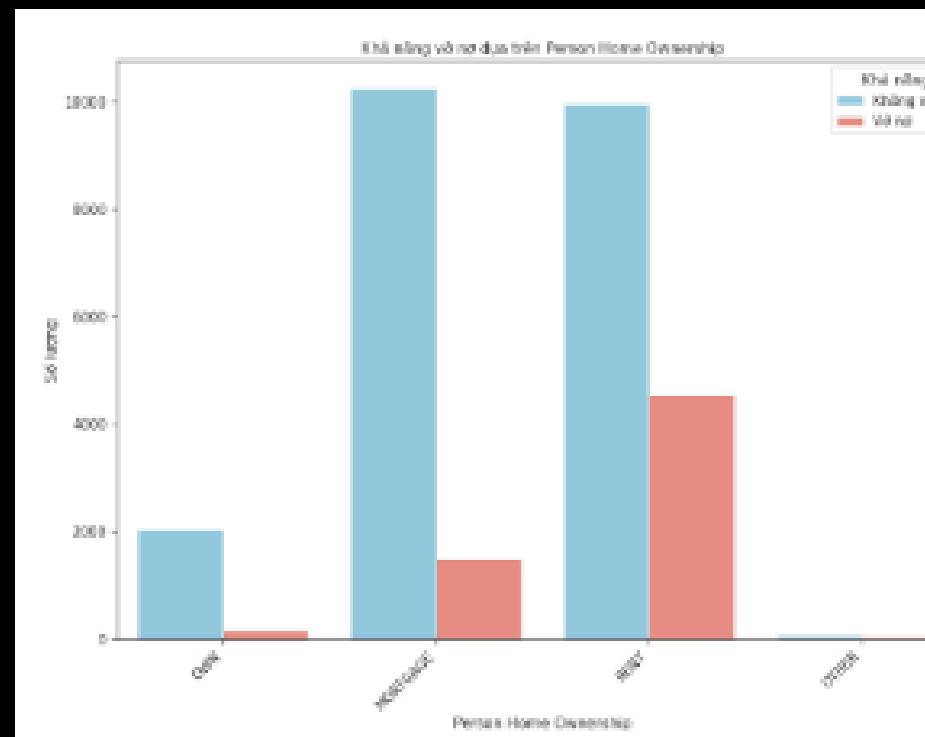
Phân phối của biến định lượng





Trực quan hóa dữ liệu

Mức độ ảnh hưởng của các biến phân loại đối với rủi ro tín dụng



Bảng số lượng thống kê trên Person Home Ownership:

Số lượng Phản trăm

person_home_ownership	Số lượng	Phản trăm
RENT	14494	50.87
MORTGAGE	11732	41.18
OWN	2174	7.63
OTHER	93	0.33

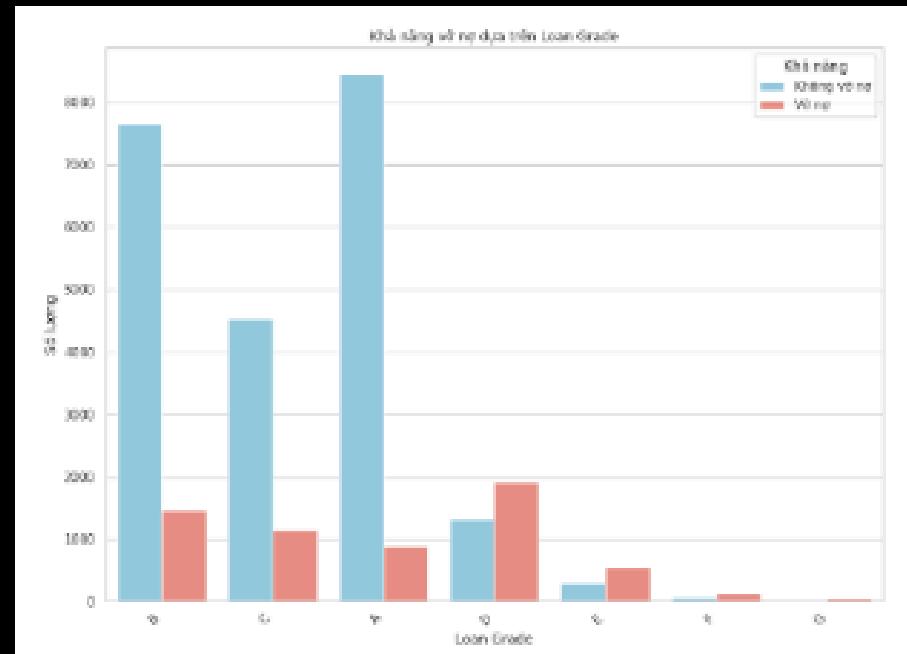
Bảng số lượng thống kê trên Loan Intent:

Số lượng Phản trăm

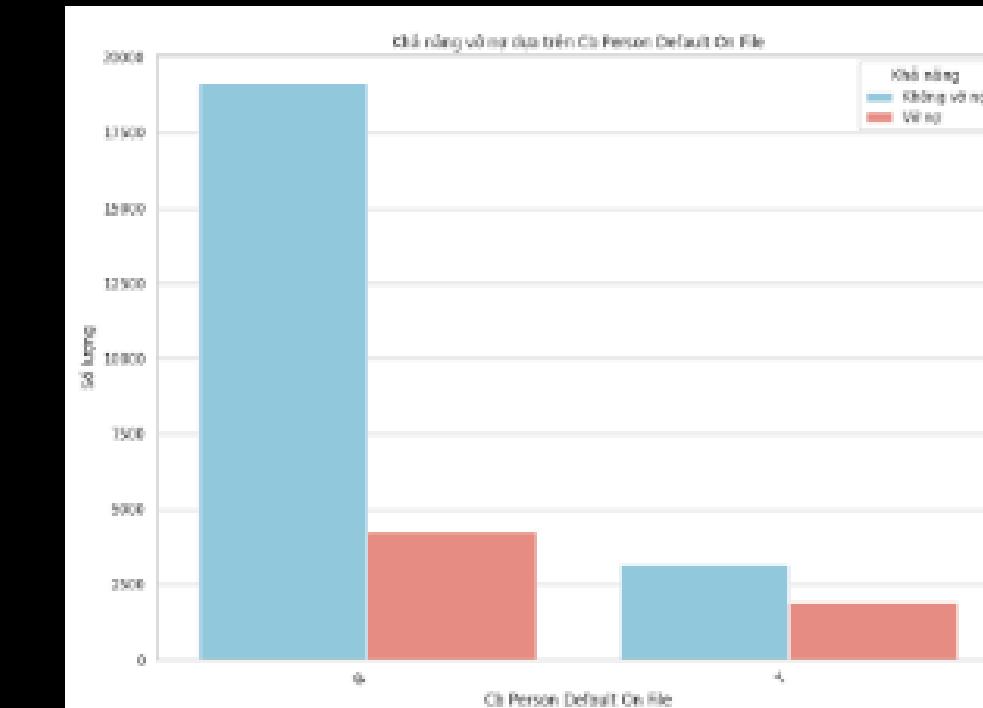
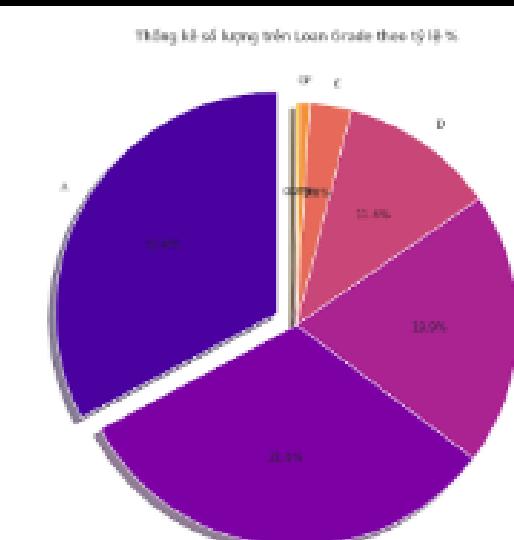
loan_intent	Số lượng	Phản trăm
EDUCATION	5668	19.89
MEDICAL	5268	18.49
VENTURE	4967	17.43
PERSONAL	4856	17.04
DEBTCONSOLIDATION	4547	15.96
HOMEIMPROVEMENT	3187	11.19

Trực quan hóa dữ liệu

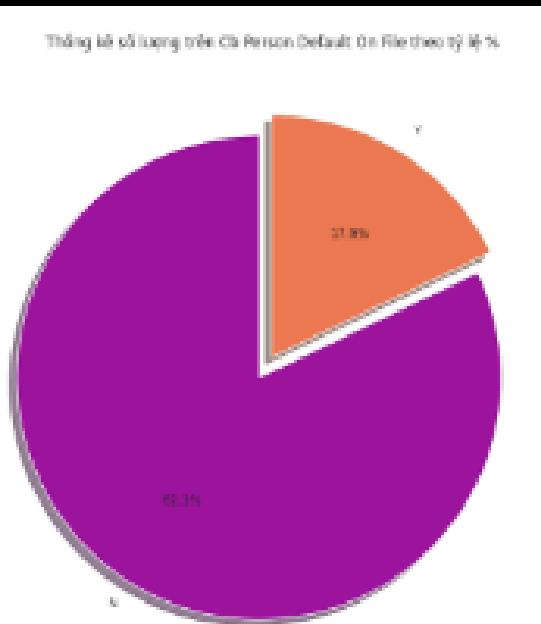
Mức độ ảnh hưởng của các biến phân loại đối với rủi ro tín dụng



Bảng số lượng thống kê trên Loan Grade:
Số lượng Phần trăm



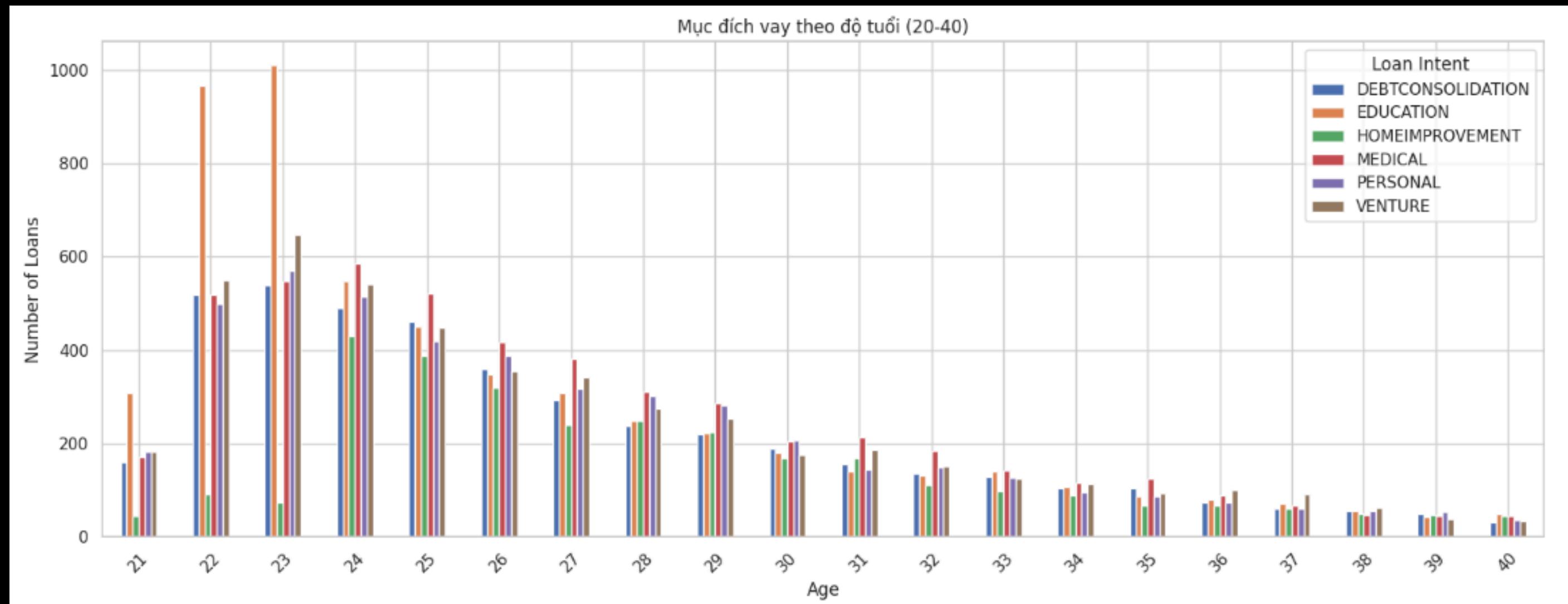
Điều này là một biểu đồ cột với hai series: 'Không vỡ nợ' (blue) và 'Vỡ nợ' (red). Trục Y là 'Số Lượng' (Quantity) từ 0 đến 20000. Trục X là 'Cb Person Default On File' (Thứ tự) N, Y.



Trực quan hóa dữ liệu

Yếu tố cản trở khi xét duyệt khoản vay dựa trên phân tích các biến định lượng

Lý do vay thay đổi theo độ tuổi

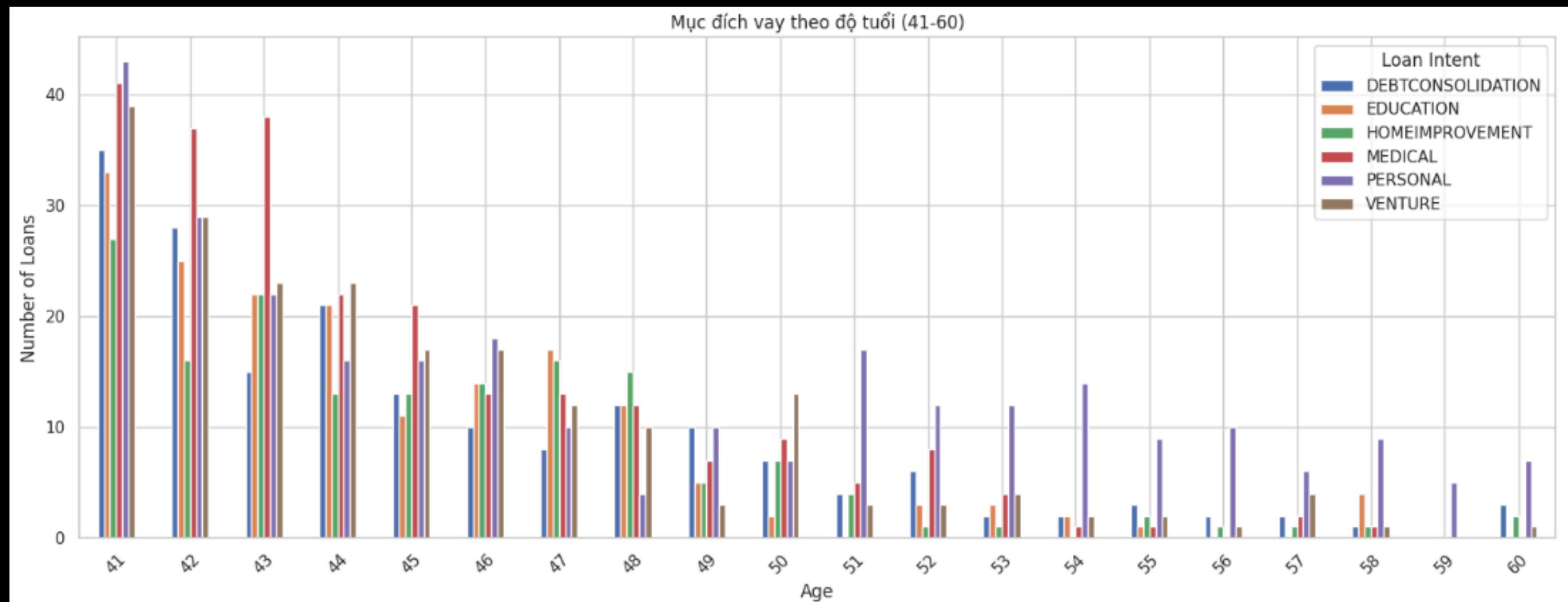


Người từ 20-40 tuổi: Mục đích vay đa dạng, tập trung vào giáo dục và cải thiện nhà cửa.
Nhu cầu vay giảm dần khi tuổi tăng, nhưng nhu cầu cho y tế và cá nhân tăng lên

Trực quan hóa dữ liệu

Yếu tố cản trở khi xét duyệt khoản vay dựa trên phân tích các biến định lượng

Lý do vay thay đổi theo độ tuổi

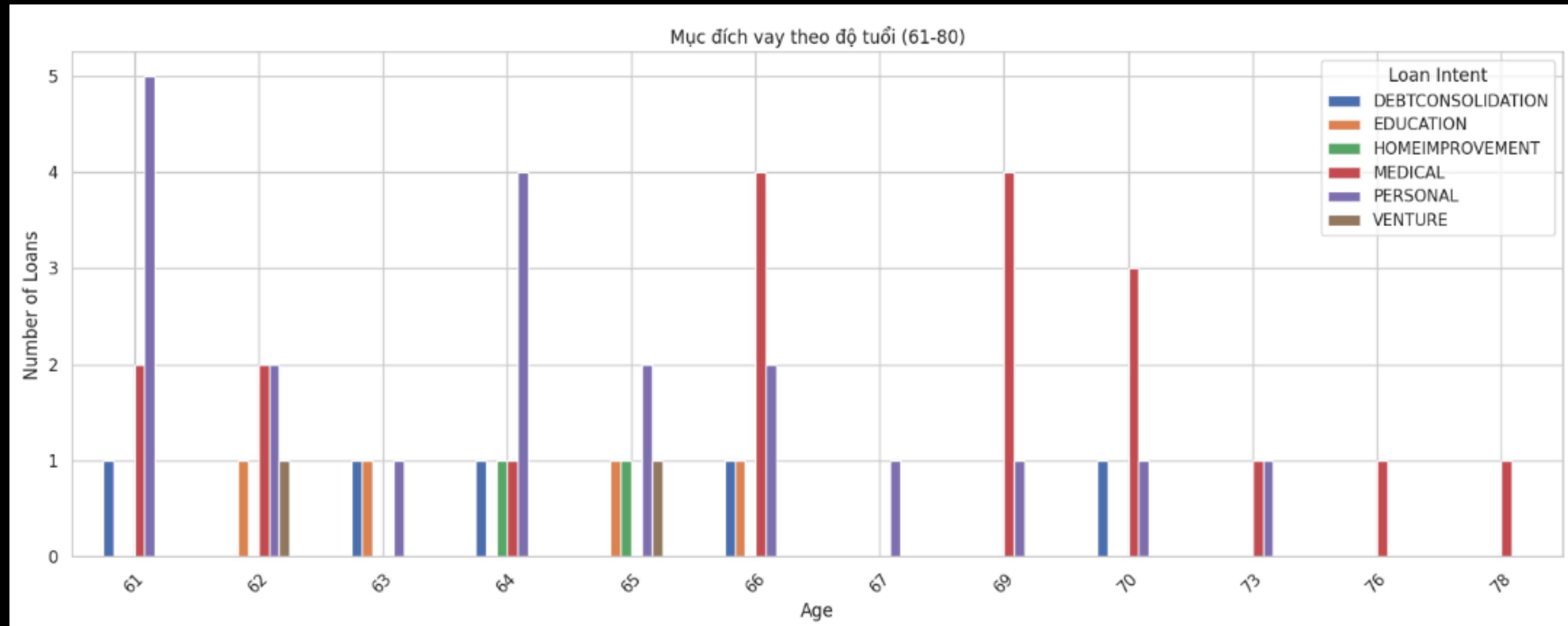


Người từ 41-60 tuổi: Hợp nhất nợ và vay cá nhân là phổ biến nhất, cho thấy sự tập trung vào quản lý tài chính. Nhu cầu vay giảm dần khi độ tuổi tăng, nhưng cải thiện nhà cửa và y tế vẫn duy trì ổn định.

Trực quan hóa dữ liệu

Yếu tố cản trở khi xét duyệt khoản vay dựa trên phân tích các biến định lượng

Lý do vay thay đổi theo độ tuổi

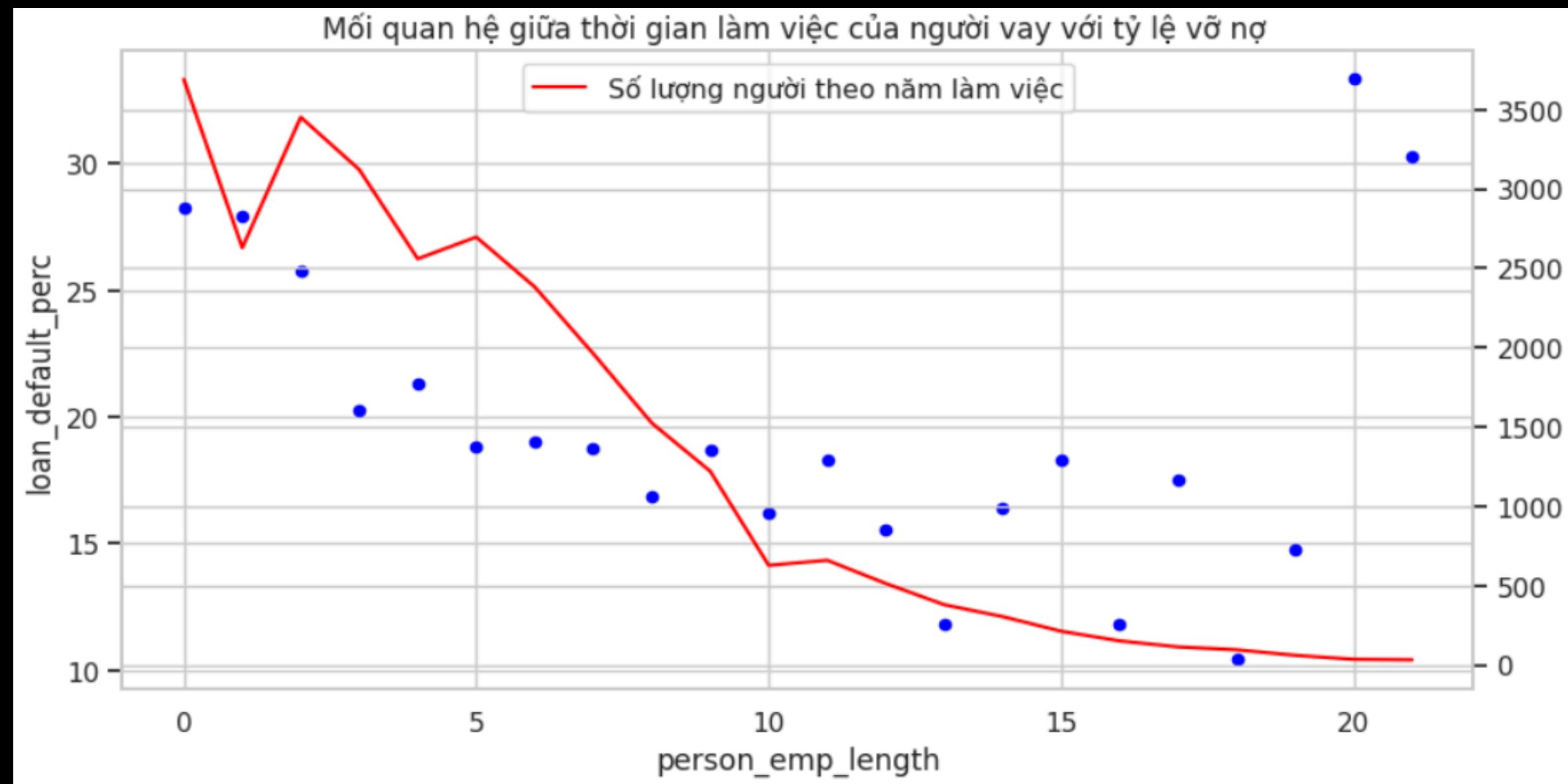


Người từ 61-80 tuổi:Nhu cầu vay giảm đáng kể ở độ tuổi này, chủ yếu vay cho mục đích y tế và cá nhân.

Trực quan hóa dữ liệu

Yếu tố cản trở khi xét duyệt khoản vay dựa trên phân tích các biến định lượng

Tác động của độ dài năm làm việc đến xu hướng gây ra vỡ nợ

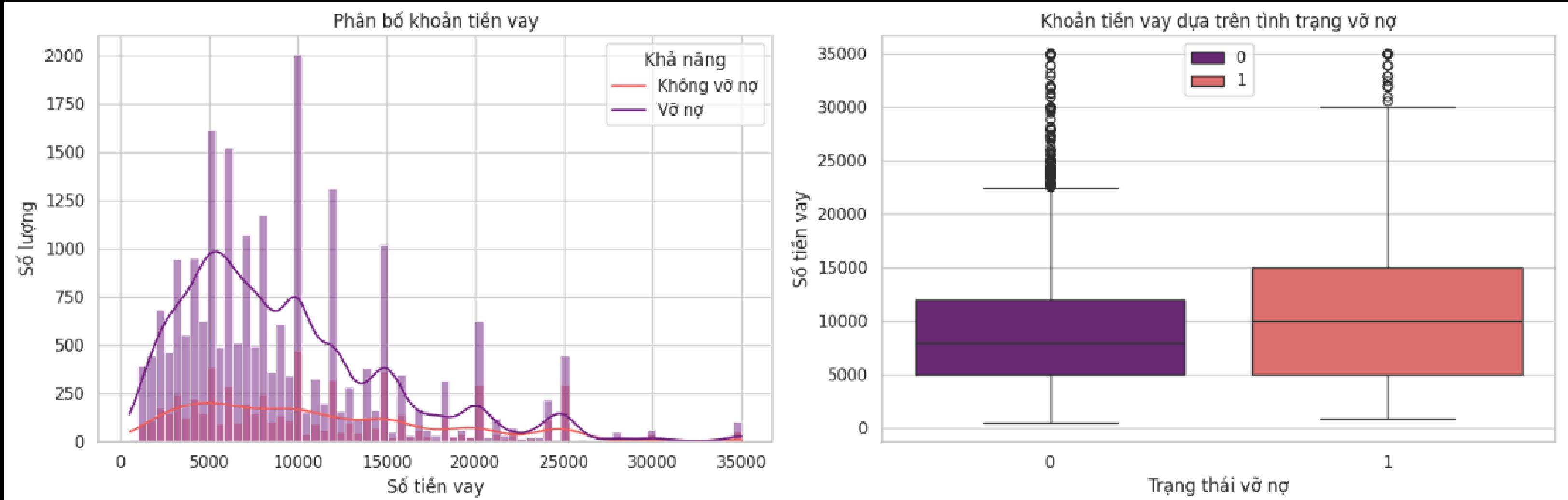


Thời gian làm việc là yếu tố quan trọng cân cân nhắc: khi tỷ lệ vỡ nợ giảm khi thời gian làm việc tăng, cho thấy thời gian làm việc dài hơn giúp giảm rủi ro vỡ nợ.

Trực quan hóa dữ liệu

Yếu tố cản trở khi xét duyệt khoản vay dựa trên phân tích các biến định lượng

Ảnh hưởng của giá trị khoản vay đến khả năng vỡ nợ

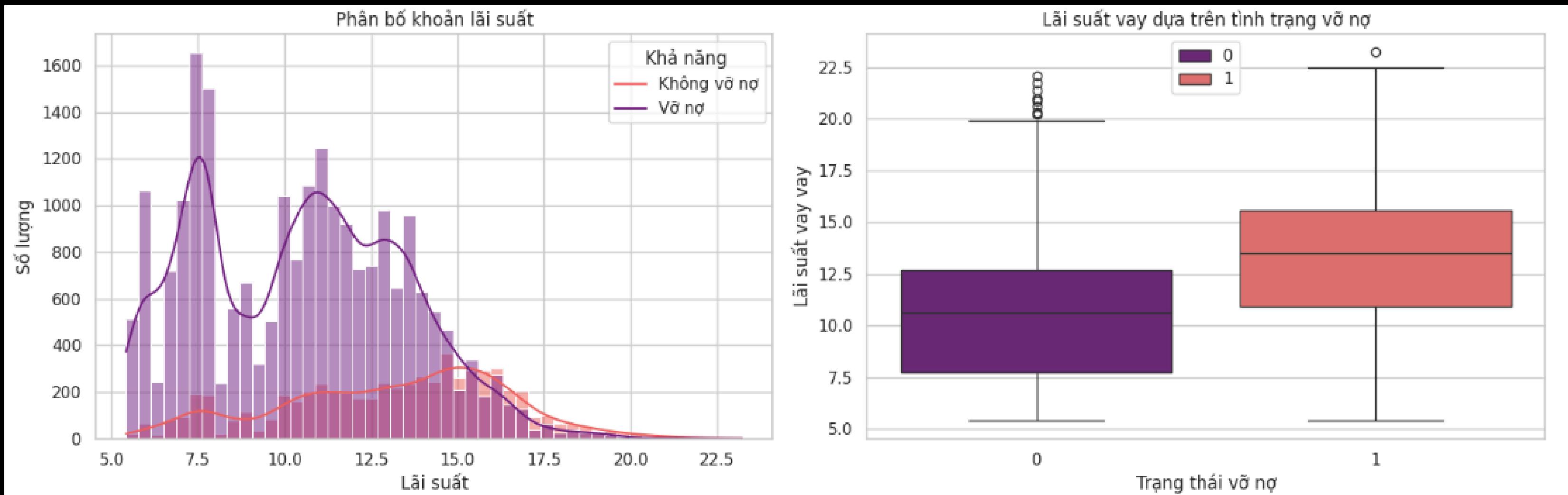


Biểu đồ cho thấy các khoản vay vỡ nợ có xu hướng tập trung ở mức khoản vay cao hơn đặc biệt là những khoản vay trên 10.000USD.

Trực quan hóa dữ liệu

Yếu tố cản trở khi xét duyệt khoản vay dựa trên phân tích các biến định lượng

Ảnh hưởng của lãi suất đến khả năng vỡ nợ trong phê duyệt khoản vay

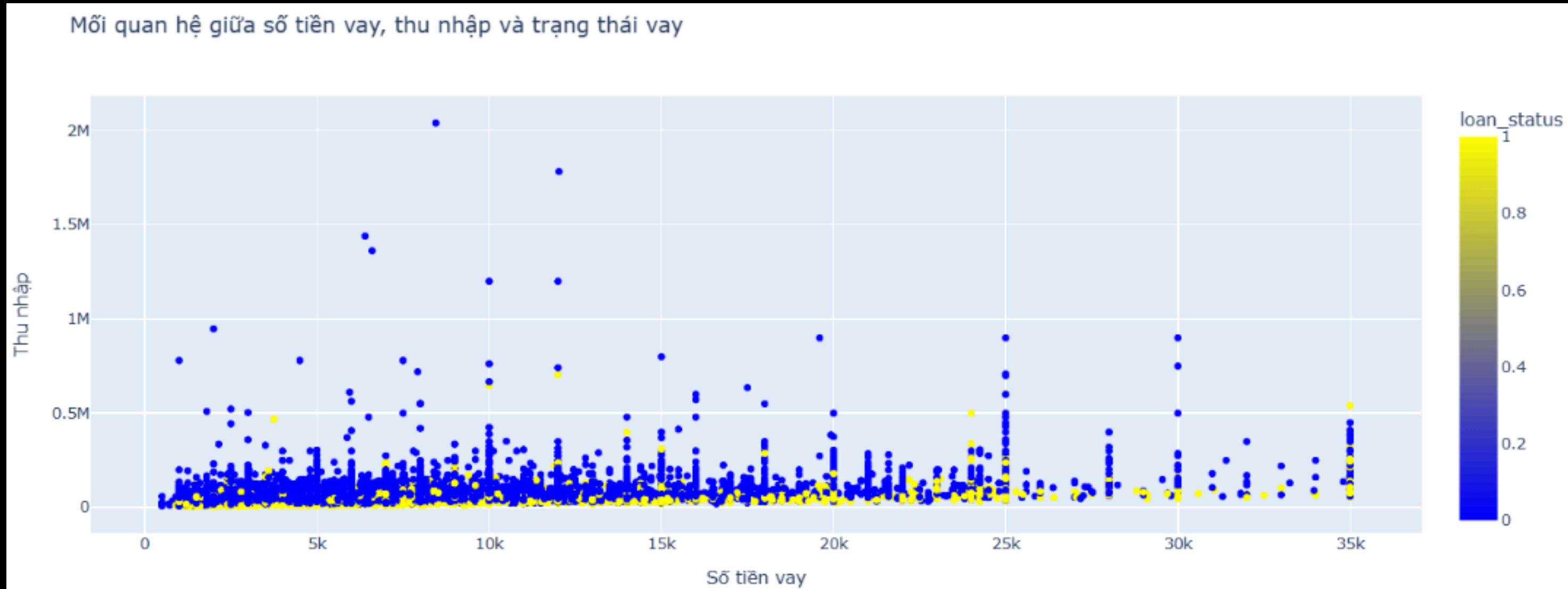


Để giảm nguy cơ vỡ nợ, lãi suất cho vay nên duy trì dưới 10%. Lãi suất này giúp giảm thiểu rủi ro vỡ nợ so với các mức lãi suất cao hơn.

Trực quan hóa dữ liệu

Yếu tố cản trở khi xét duyệt khoản vay dựa trên phân tích các biến định lượng

Mối liên hệ giữa thu nhập và mức tiền vay đối với rủi ro vỡ nợ

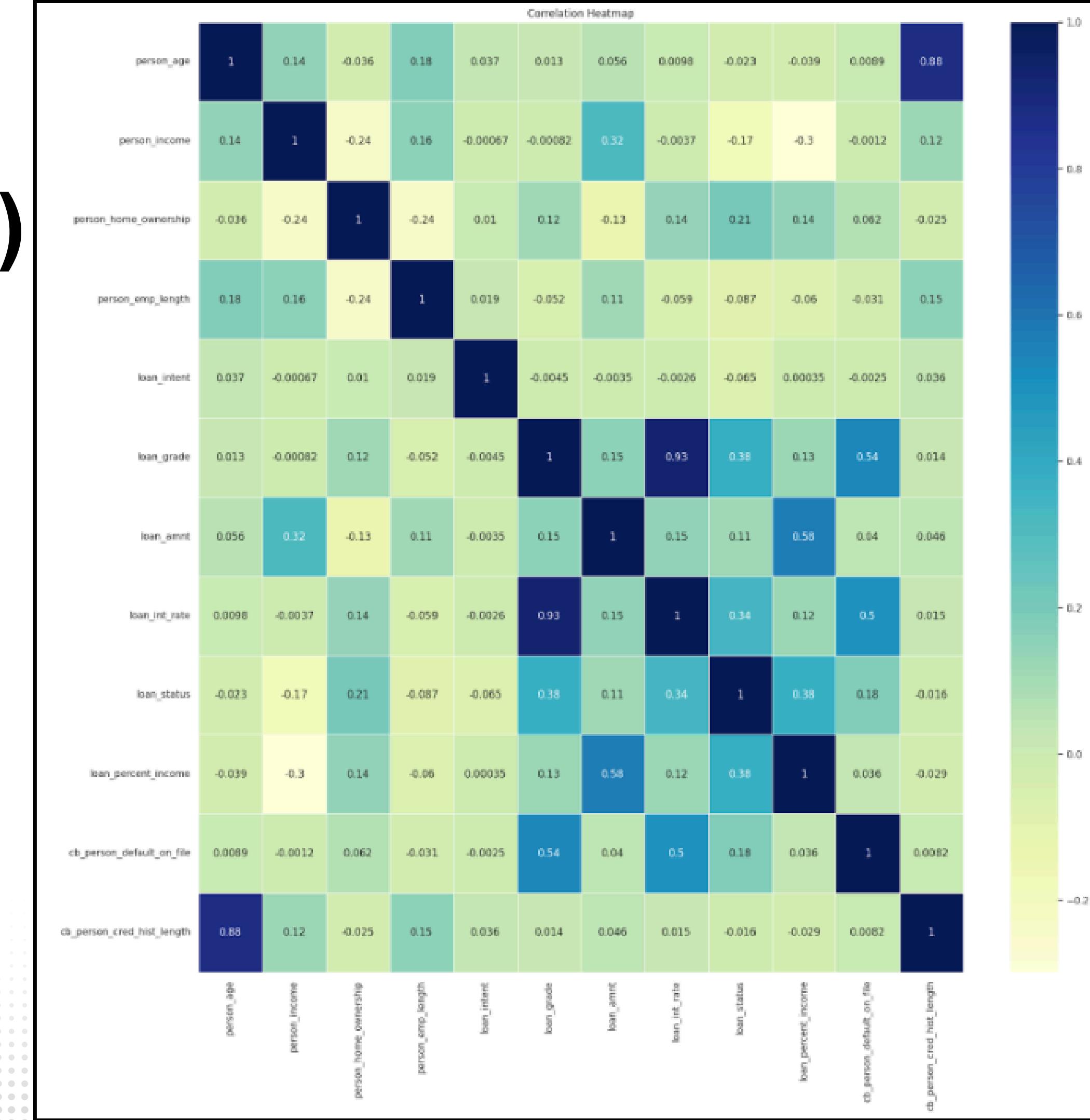


Phần lớn các khoản vay có giá trị dưới 25,000 USD, và thu nhập cá nhân tập trung dưới 500,000 USD. Trung bình các khoản vay vỡ nợ cao hơn, đặc biệt là ở mức trên 15,000 USD. Những người có thu nhập trên 500,000 USD ít có dấu hiệu vỡ nợ hơn, dù vay các khoản lớn.

Correlation (Ma trận tương quan)

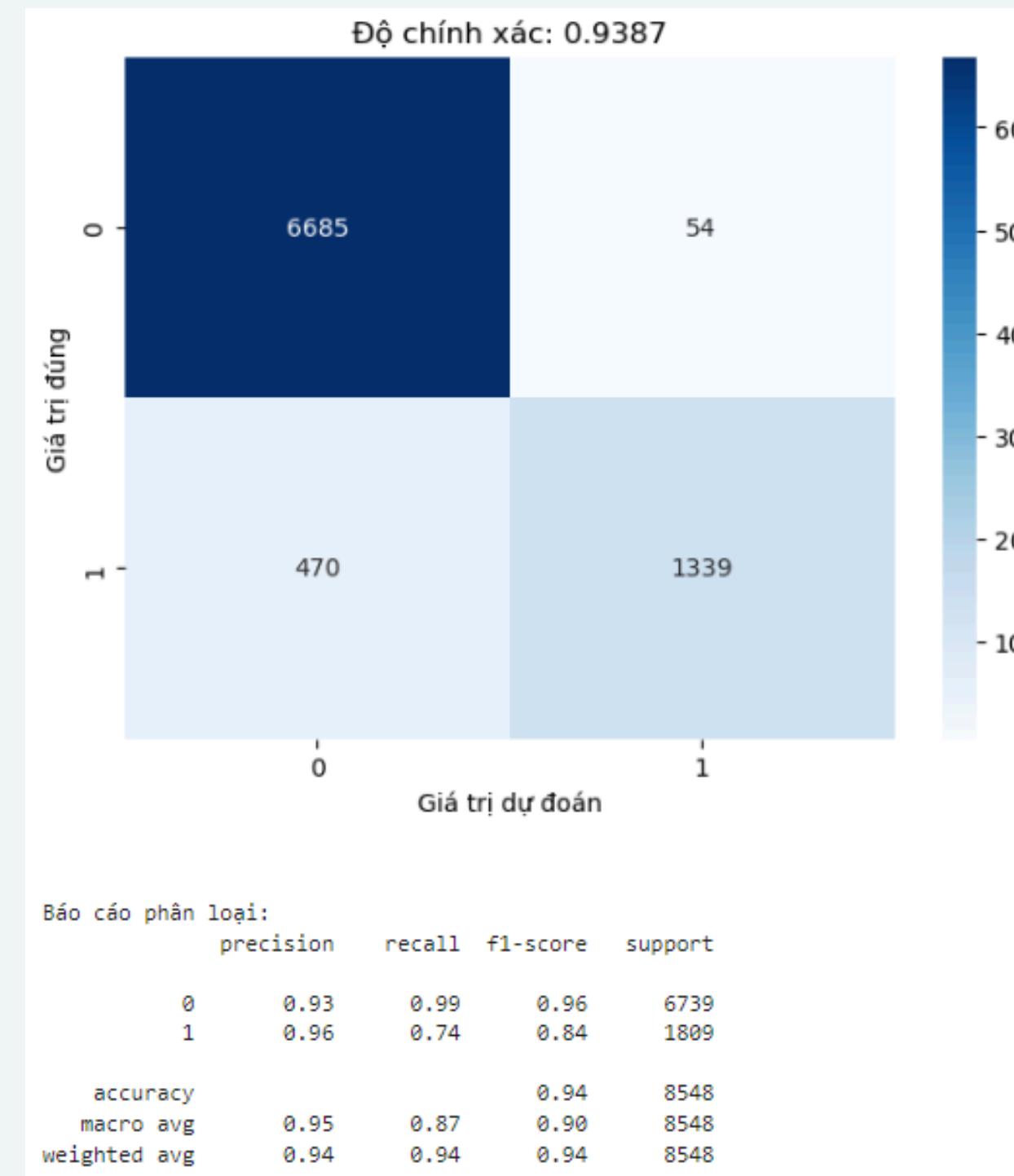


Các thuộc tính trong ma trận tương quan cung cấp thông tin quan trọng, bổ sung cho nhau trong dự đoán rủi ro tín dụng. Việc sử dụng đầy đủ 11 thuộc tính (trừ loan_status) giúp mô hình toàn diện hơn. Các thuật toán như CatBoost, Decision Tree, XGBoost, và Random Forest xử lý tốt mối quan hệ phức tạp, cải thiện độ chính xác và hiệu suất dự đoán.



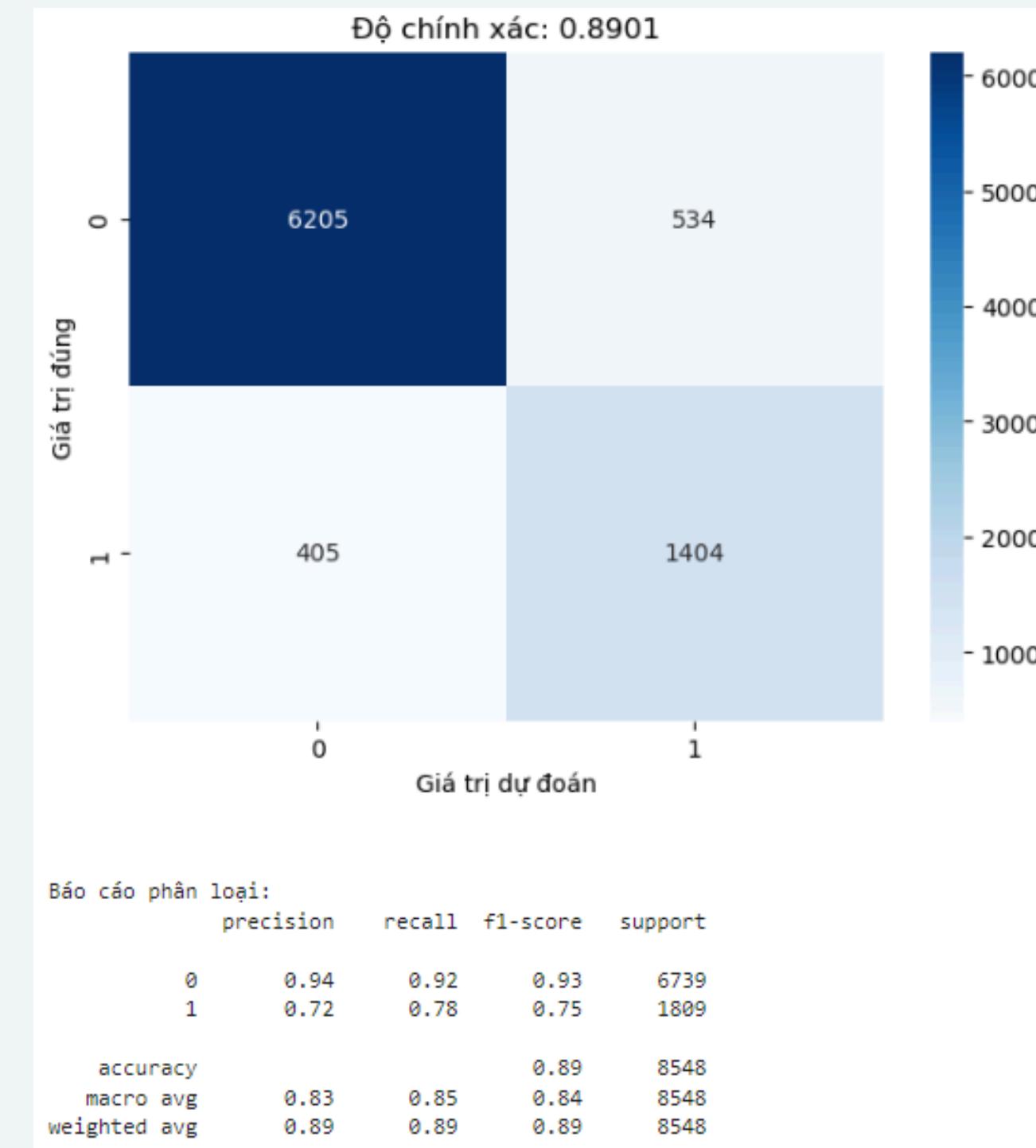
Mô hình thuật toán

CatBoost



Chênh lệch score giữa Train (0.9573) và Test (0.9387) là 2%, với độ chính xác 93.87%, cho thấy mô hình hiệu quả.

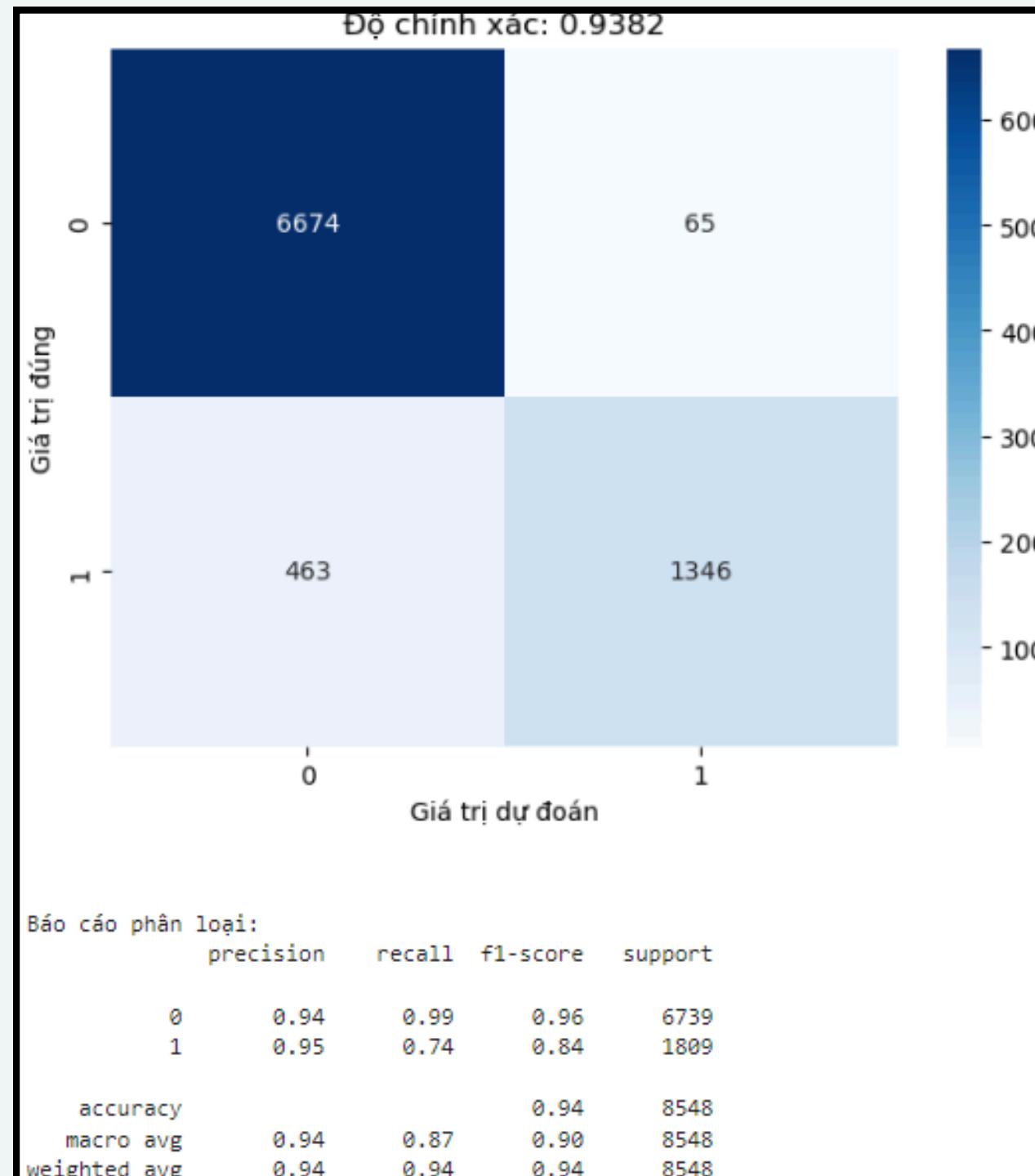
Decision Tree



Chênh lệch score giữa Train (1.0000) và Test (0.8901) là 11%, thể hiện overfitting, giảm độ tin cậy khi áp dụng thực tế.

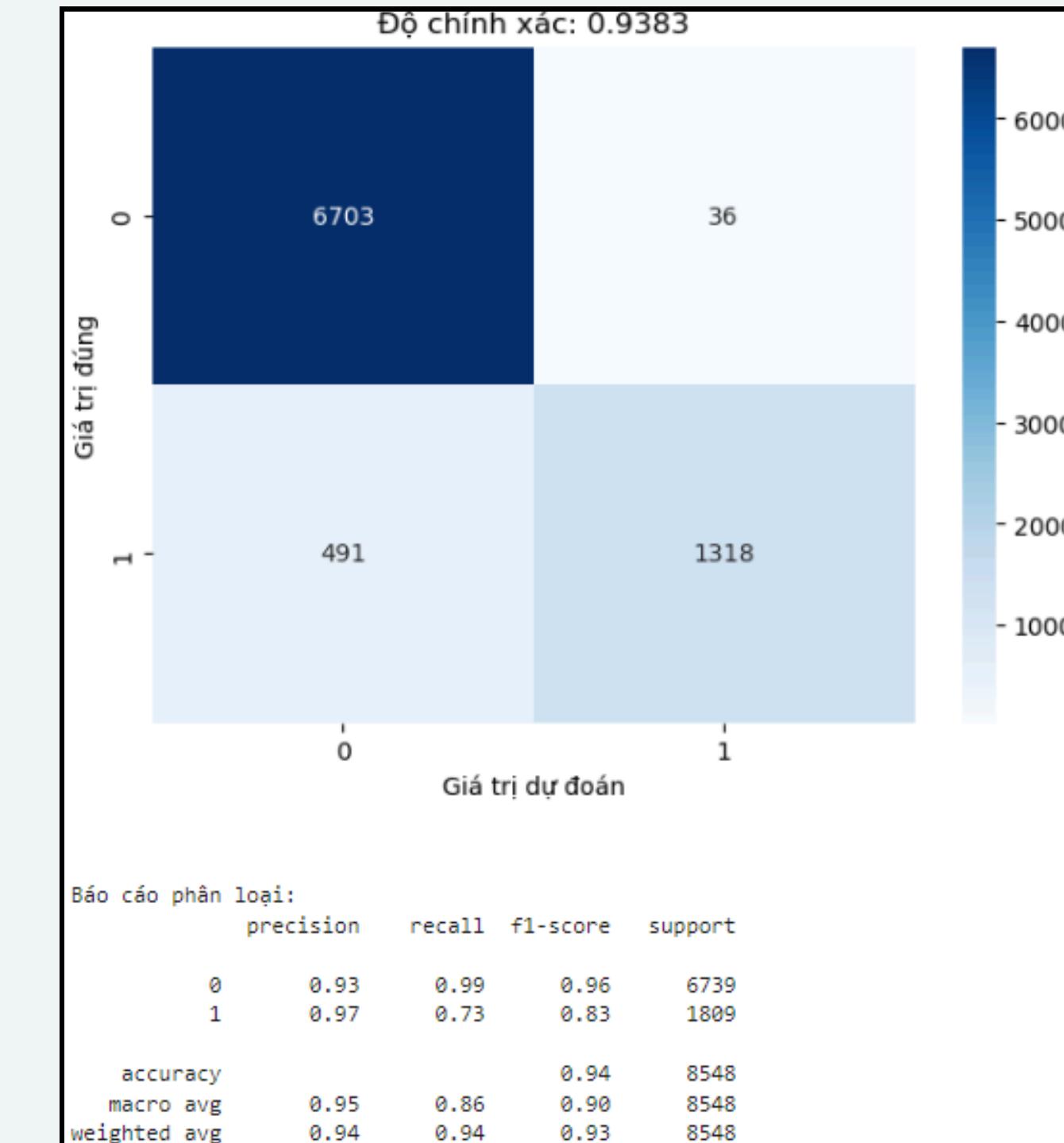
Mô hình thuật toán

XGBboost



Chênh lệch score giữa Train (0.9601) và Test (0.9382) là 2%, với độ chính xác 93.82%, mô hình phù hợp để phân tích.

Random Forest



Chênh lệch score giữa Train (1.0000) và Test (0.9383) là 6%, cho thấy overfitting, nhưng độ chính xác tổng thể (93.83%) vẫn cao.

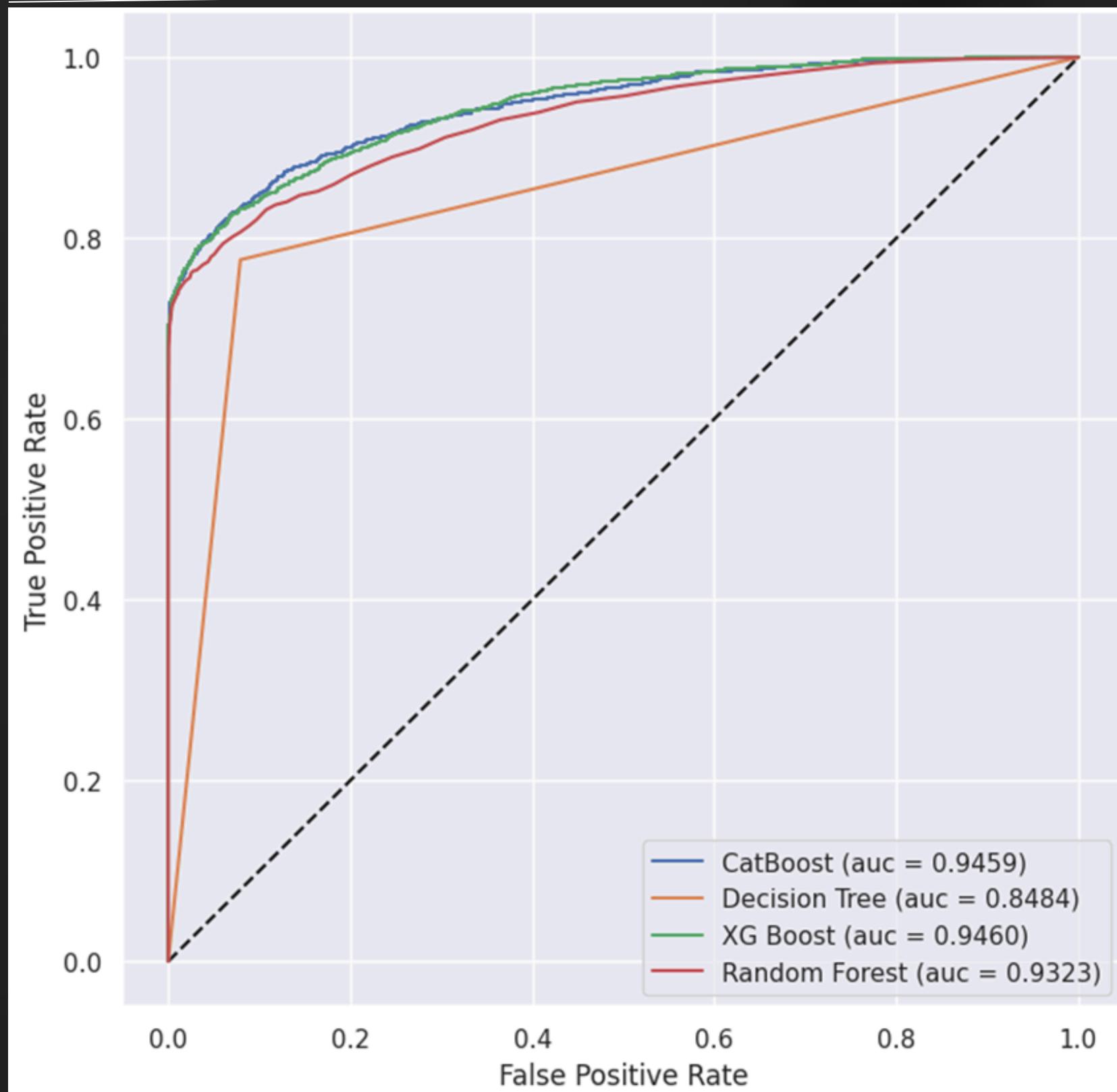
Đánh giá mô hình

So sánh các chỉ số đánh giá hiệu suất của các mô hình

	Độ chính xác (Accuracy)	Độ chuẩn xác(Precision)	Độ nhạy(Recall)	Điểm F1(F1 Score)
CatBoost	0.9387	0.9612	0.7402	0.8364
Decision Tree	0.8901	0.7245	0.7761	0.7494
XgBoost	0.9382	0.9539	0.7441	0.8360
Random Forest	0.9383	0.9734	0.7286	0.8334

Đánh giá mô hình

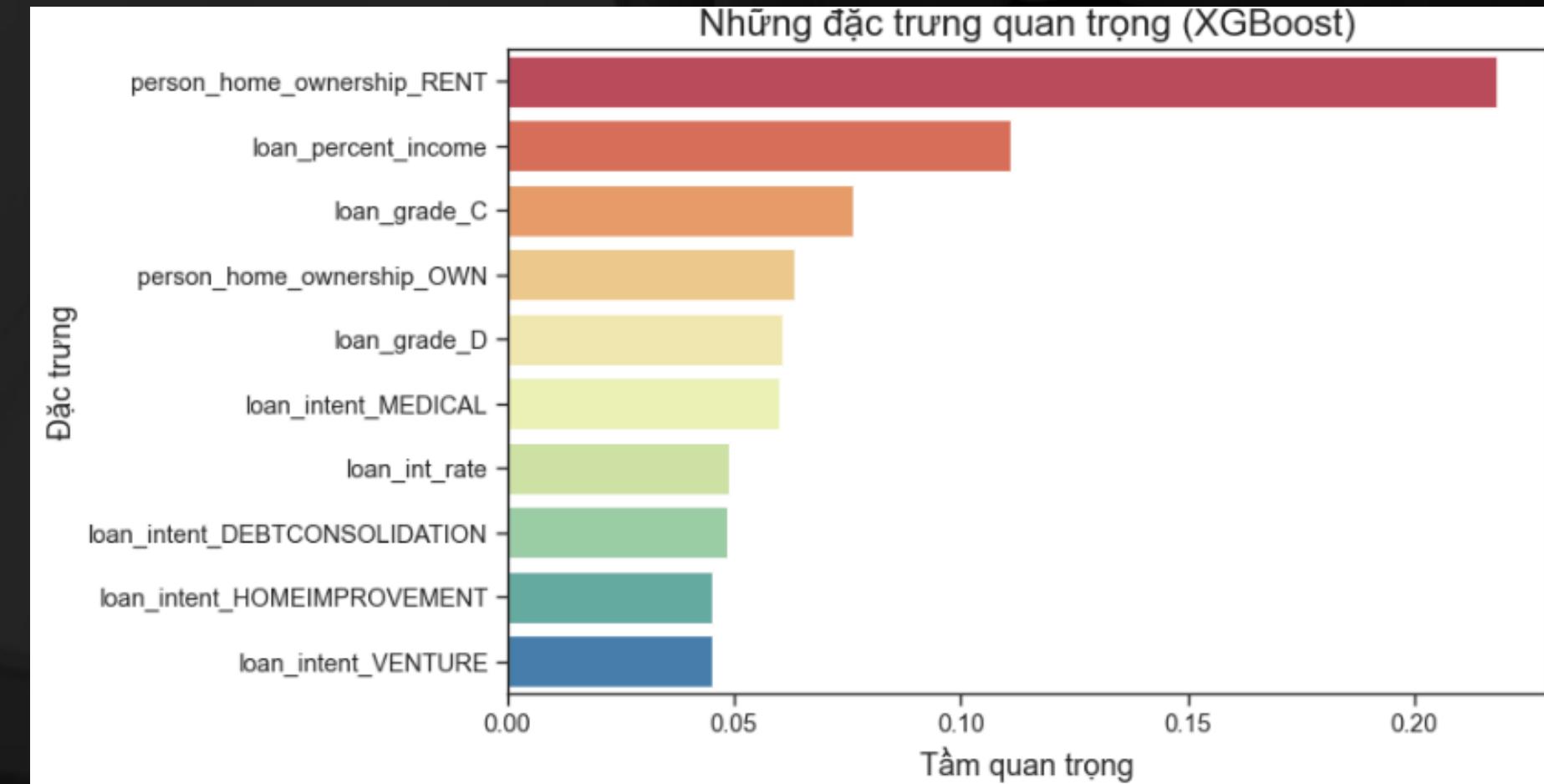
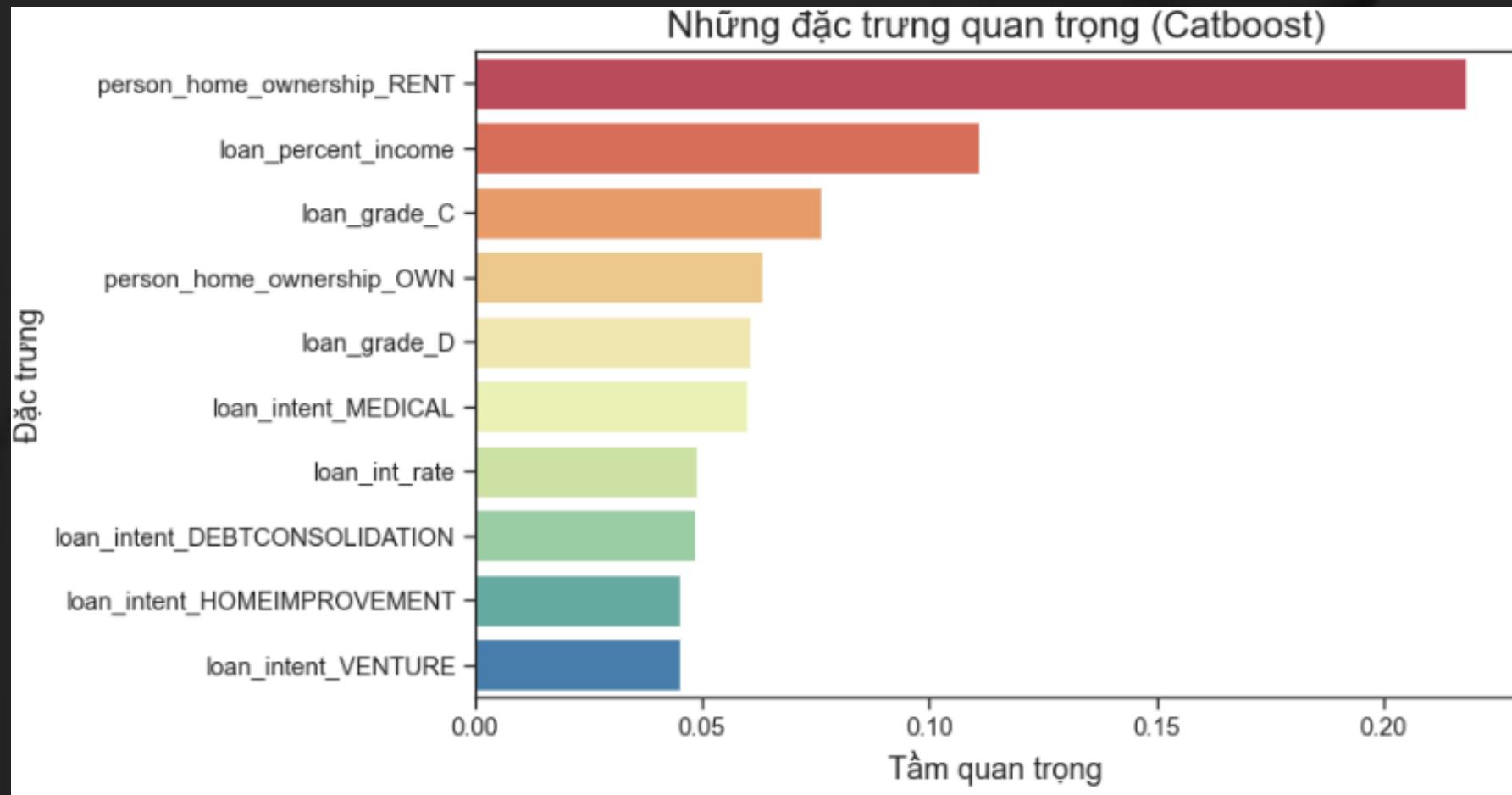
Đánh giá hiệu quả các mô hình qua giá trị AUC của ROC



XGBoost và CatBoost là hai mô hình tối ưu nhất cho bài toán dự đoán rủi ro tín dụng nhờ giá trị AUC cao có thể phân biệt giữa các lớp một cách hiệu quả (rủi ro vỡ nợ và không rủi ro).

Đánh giá mô hình

Đánh giá những đặc trưng quan trọng trong mô hình dự đoán



- **CatBoost** ưu tiên các đặc trưng liên quan đến tỷ lệ thu nhập và khoản vay hơn, trong khi **XGBoost** có xu hướng nhấn mạnh quyền sở hữu nhà và các biến liên quan đến mục đích vay.
- Mức độ phân tách tầm quan trọng giữa các đặc trưng của CatBoost rõ ràng hơn so với XGBoost, thể hiện qua việc một số đặc trưng chiếm ưu thế lớn.

Giao diện xây dựng trên Streamlit:

Dự Đoán Rủi Ro Tín Dụng 

Nhập thông tin của khách hàng để dự đoán khoản vay có khả năng vỡ nợ hay không

Tuổi
20

Thu nhập hàng năm (\$)
50000

Số tiền vay (\$)
10000

Số năm làm việc
5

Lãi suất khoản vay (%)
5.00

Tỷ lệ thu nhập trên số tiền vay: 20.00%

Số năm lịch sử tín dụng
10

Sở hữu nhà
MORTGAGE

Mục đích vay
DEBTCONSOLIDATION

Điểm tín dụng
A

Lịch sử vỡ nợ
N

Dự Đoán

Xây dựng mô hình dự đoán rủi ro tín dụng trên Streamlit

Nhập dữ liệu

Tuổi
28

Thu nhập hàng năm (\$)
25000

Số tiền vay (\$)
5850

Số năm làm việc
1

Lãi suất khoản vay [%]
20.10

Tỷ lệ thu nhập trên số tiền vay: 23.40%

Số năm lịch sử tín dụng
1

Sở hữu nhà
OTHER

Mục đích vay
VENTURE

Điểm tín dụng
C

Lịch sử vỡ nợ
Y

Dự Đoán

Khách hàng nằm trong nhóm nguy cơ cao vỡ nợ khoản vay (Xác suất: 70.22%)

Tuổi
28

Thu nhập hàng năm (\$)
25000

Số tiền vay (\$)
5850

Số năm làm việc
1

Lãi suất khoản vay [%]
12.06

Tỷ lệ thu nhập trên số tiền vay: 23.40%

Số năm lịch sử tín dụng
3

Sở hữu nhà
OWN

Mục đích vay
EDUCATION

Điểm tín dụng
C

Lịch sử vỡ nợ
N

Dự Đoán

Khách hàng có nguy cơ thấp vỡ nợ khoản vay (Xác suất: 98.56%)

Kết luận

Phần trực quan dữ liệu:

Phân tích các yếu tố định lượng và phân loại đã làm rõ nhiều khía cạnh quan trọng trong việc xét duyệt tín dụng. Một số điểm nổi bật gồm:

- Tập trung vào các yếu tố như giá trị khoản vay, tỷ lệ vay theo tổng thu nhập, và lãi suất để giảm thiểu rủi ro vỡ nợ.
- Cân nhắc thời gian làm việc và lịch sử tín dụng khi đánh giá hồ sơ vay.
- Cần chú ý trong việc xây dựng các mô hình dự báo chính xác hơn bằng cách xử lý dữ liệu mất cân bằng và đa cộng tuyến từ các biến có tương quan mạnh trong tập dữ liệu.

Tóm lại, phân tích này cung cấp cơ sở cho việc ra xây dựng chiến lược xét duyệt khoản vay hiệu quả hơn, đồng thời hỗ trợ trong việc xây dựng mô hình dự đoán rủi ro và nâng cao hiệu quả xét duyệt tín dụng.

Kết luận

Phần machine learning:

Trong dự án này, một số kết quả quan trọng được rút ra như sau:

- Đề tài sử dụng các mô hình Machine Learning để tìm ra mô hình tối ưu cho việc dự đoán rủi ro tín dụng (Credit Risk). Mô hình hiệu quả nhất được xác định là CatBoost.
- Độ chính xác của mô hình này đạt 93,87%.
- Bên cạnh đó, ROC của mô hình CatBoost rất cao ($auc = 95\%$).
- XGBoost cũng là một mô hình cho kết quả dự đoán tương đối tốt và không kém gì CatBoost trong dự án này.

Tóm lại, việc xây dựng mô hình dự đoán rủi ro tín dụng là rất quan trọng đối với các ngân hàng và tổ chức tín dụng. Vì vậy, khi thực hiện dự đoán rủi ro tín dụng, việc lựa chọn phương pháp phù hợp là rất cần thiết, nhằm cung cấp thông tin chính xác cho các dự báo, yêu cầu người thực hiện phải hiểu rõ cách áp dụng các phương pháp này một cách hiệu quả.

Xin cảm ơn!

Vui lòng liên hệ với chúng tôi nếu bạn có bất kỳ câu hỏi nào.

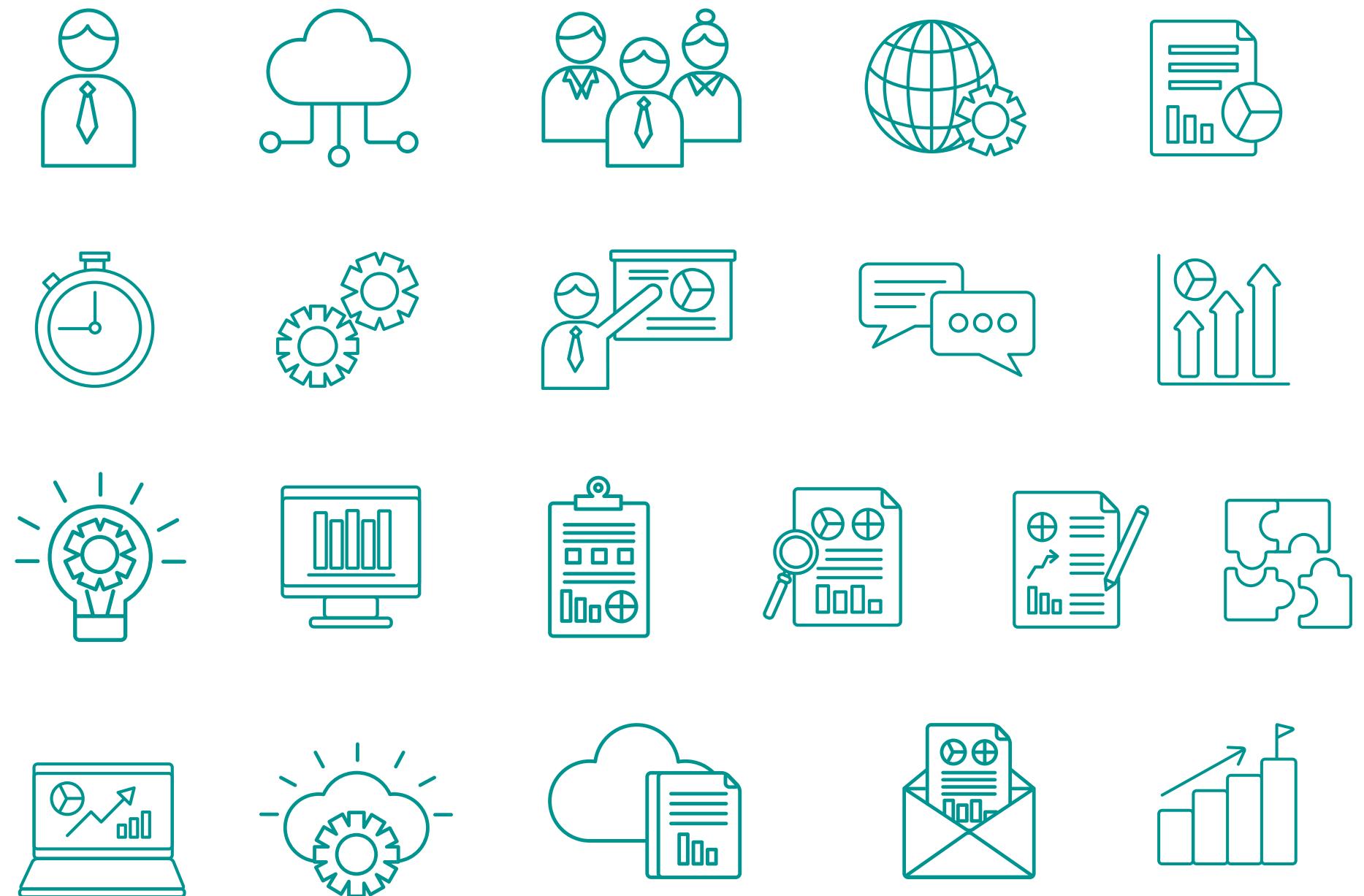
Số điện thoại
0912 3456 7890

Địa chỉ E-mail
xinchao@trangwebhay.vn



Trang tài nguyên

Hãy sử dụng những biểu tượng và minh họa này trong Bản Thuyết Trình Canva của bạn. Thiết kế vui vẻ!



Trang tài nguyên

Tìm điều kỳ diệu và thú vị khi trình bày với Bản Thuyết Trình Canva. Nhấn các phím sau khi ở chế độ Trình bày!

B để tạo hiệu ứng mờ

C để tạo hoa giấy

D để tạo tiếng trống

O để tạo bong bóng

Q để tắt tiếng

X để đóng

Bất kỳ số nào từ 0-9 để hẹn giờ