Đánh giá điểm tín dụng khách hàng - Credit Scoring

Nhận các hồ sơ vay vốn, đánh giá rủi ro tín dụng, ra quyết định về việc cho vay và giám sát việc hoàn trả vốn gốc và lãi. Đánh giá liệu một ứng viên mới có đủ tin cậy để được vay hay không nếu các chỉ tiêu đặc trưng của người nộp đơn được cung cấp, như thu nhập, tình trạng hôn nhân, tuổi, lịch sử tín dụng (chẳng hạn đã từng nợ xấu hay chưa), v.v.

Đầu vào: Hồ sơ khách hàng, lịch sử tín dụng, thông tin giao dịch, tài sản, … của khách hàng.

Đầu ra: Điểm tín dụng - xác suất vỡ nợ của khách hàng.

Dự đoán tình trạng vay vốn là vỡ nợ hay không vỡ nợ dựa trên bộ dữ liệu **credit risk dataset** (kaggle dataset).

1. Exploratory Data Analysis - Phân tích khám phá dữ liệu

1.1. Kích thước dữ liệu

Bộ dữ liệu bao gồm 32,581 mẫu dữ liệu (bản ghi) và 12 trường dữ liệu

1.2. Ý nghĩa các trường dữ liệu

Bảng: mô tả các trường dữ liệu

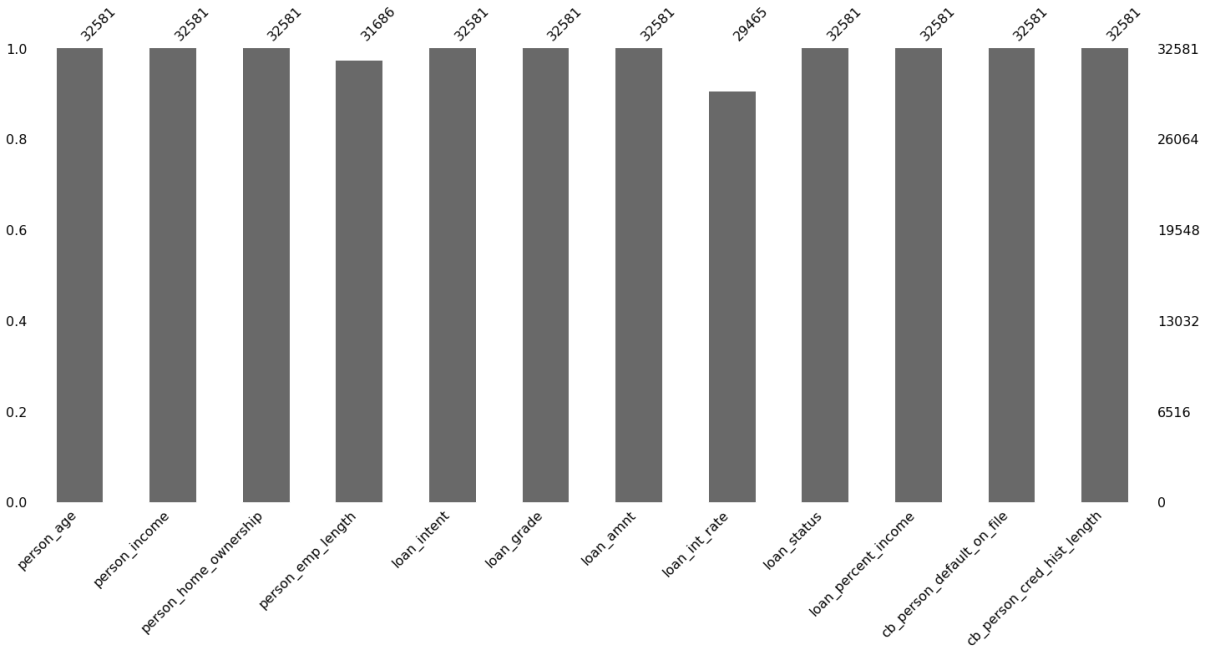
Column Name	Description	Data Type
person_age	tuổi của khách hàng	numerical - int
person_income	thu nhập hàng năm	numerical - int
person_home_ownership	phương thức sở hữu nhà ở	categorical
person_emp_length	số năm làm thuê	numerical - int
loan_intent	mục đích vay vốn	categorical
loan_amnt	khoản vay	numerical - int
loan_grade	xếp hạng cho vay	categorical
loan_int_rate	lãi suất	numerical - double
loan_percent_income	tỷ lệ khoản vay trên thu nhập	numerical - double
cb_person_default_on_file	lịch sử vỡ nợ	categorical
cb_person_cred_hist_length	số năm mở tín dụng	numerical- int
loan_status (target label)	tình trạng vay vốn (vỡ nợ / không vỡ nợ)	categorical (0/1)

person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt
22	59000	RENT	123.0	PERSONAL	D	35000
21	9600	OWN	5.0	EDUCATION	B	1000
25	9600	MORTGAGE	1.0	MEDICAL	C	5500
23	65500	RENT	4.0	MEDICAL	C	35000
24	54400	RENT	8.0	MEDICAL	C	35000

loan_int_rate	loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
16.02	1	0.59	Y	3
11.14	0	0.10	N	2
12.87	1	0.57	N	3
15.23	1	0.53	N	2
14.27	1	0.55	Y	4

Hình: một số mẫu dữ liệu

1.3 Kiểm tra trường dữ liệu bị khuyết và dữ liệu trùng lặp



Hai trường dữ liệu bị thiếu là số năm làm thuê (person_emp_length) (missing 2,7%) và lãi suất cho vay (loan_int_rate) (missing 9.5 %). Số lượng khuyết không nhiều có thể thay thế bằng giá trị trung bình hoặc trung vị trong từng trường.
Dữ liệu có khoảng 165 mẫu dữ liệu trùng lặp.

1.4. Phân phối của từng trường dữ liệu

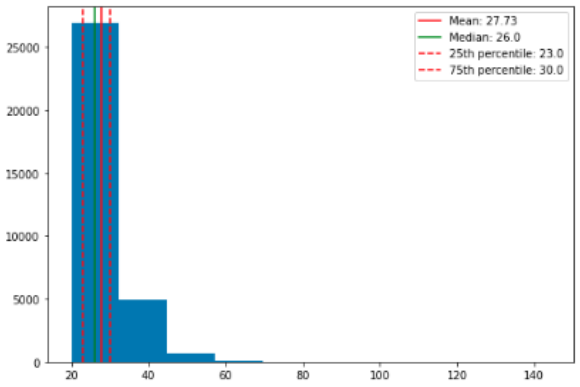
Các trường dạng số bao gồm *person_age, person_income, person_emp_length, loan_amnt, loan_int_rate, loan_percent_income, cb_person_cred_hist_length.*

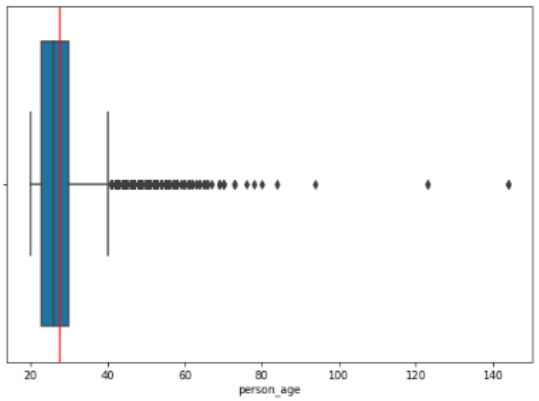
Các trường dạng hạng mục bao gồm *person_home_ownership, loan_intent, loan_grade, loan_status, cb_person_default_on_file.*

Trong đó trường *loan_status* là cột nhãn dự đoán.

Các trường dạng số:

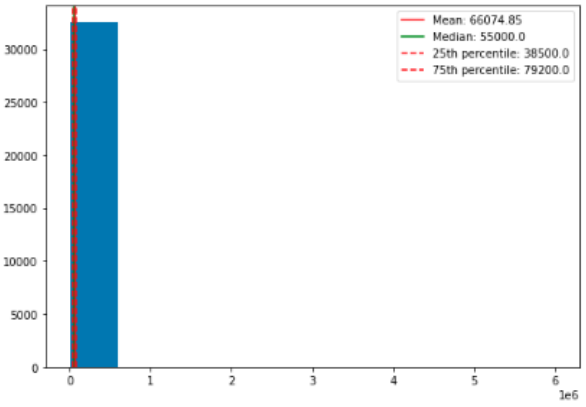
- *person_age* - tuổi của khách hàng

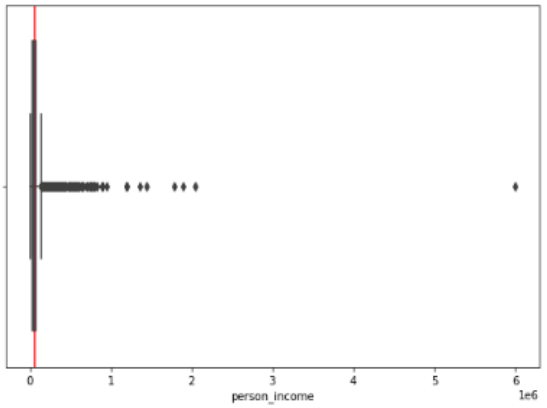




Nhận xét: Từ histogram cho thấy dữ liệu phân bố hầu hết trong khoảng tuổi từ 23 đến 30. Đồ thị có xu hướng lệch phải (right skewed), với 75% khách hàng trong khoảng tuổi thấp hơn 30. Từ boxplot thấy có khá nhiều điểm được coi là nhiễu ngoại lệ (outliers) khác biệt so với phân bố. Đặc biệt có giá trị tuổi là hơn 140, độ tuổi này cũng có thể xảy ra nhưng không hợp lý khi tham gia vay vốn.

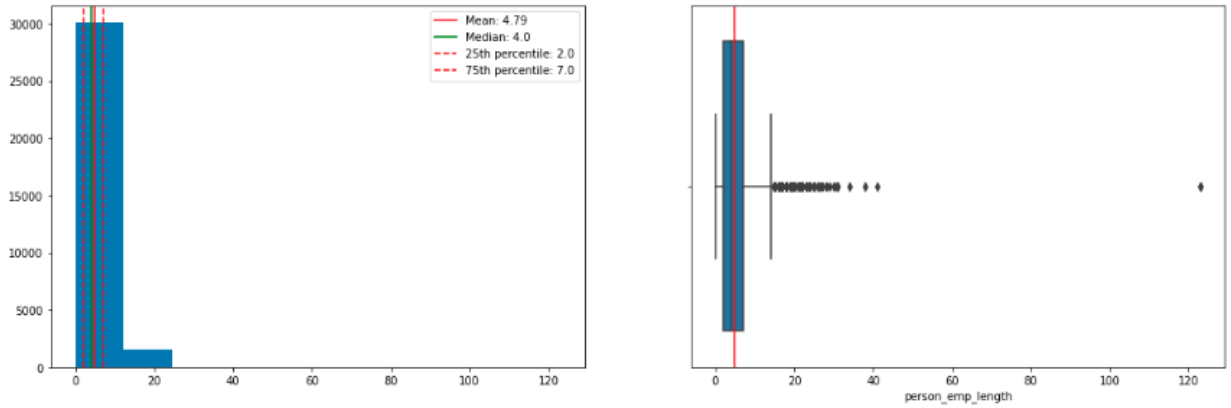
- *person_income* - thu nhập hàng năm





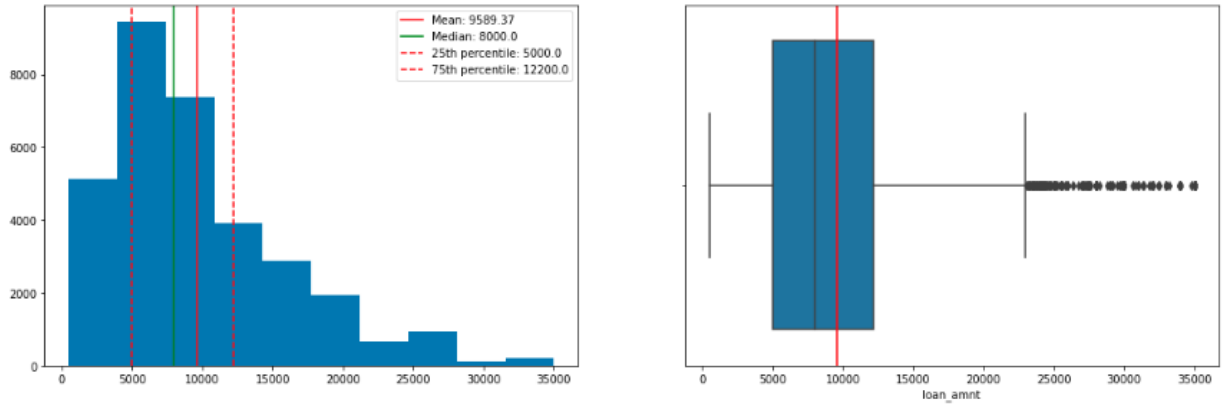
Nhận xét: Dữ liệu phân bố hầu hết trong khoảng thu nhập từ 38,500 đến 79,200. Phân bố dữ liệu lệch về bên phải, nhiều điểm nhiễu nằm xa so với phân bố tập trung của dữ liệu.

- *person_emp_length* - số năm làm thuê



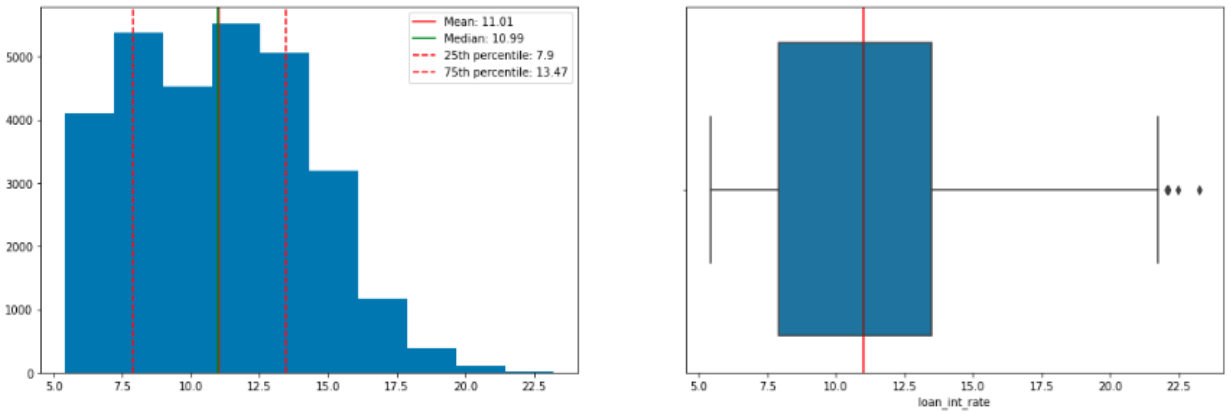
Nhận xét: Đồ thị lệch phải và có khả năng có nhiễu ngoại lệ. Với phân bố hầu hết trong khoảng từ 2 đến 7 năm. Với một mẫu dị thường có số năm làm thuê lớn nhất là 123. Có khả năng do lỗi nhập liệu thông tin.

- *loan_amnt* - khoản vay vốn



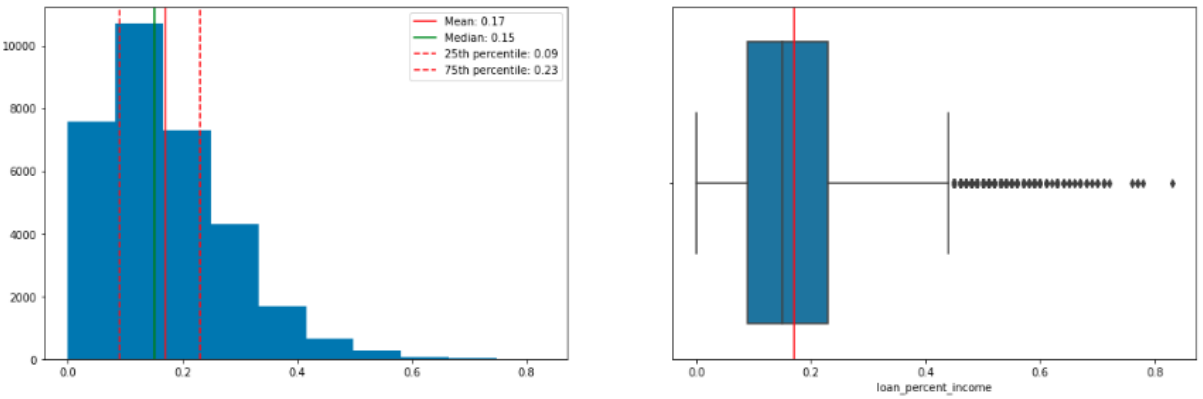
Nhận xét: Dữ liệu có xu hướng bị lệch về phải nhẹ. Khoản vay vốn tập trung nhiều trong khoảng 5000 đến 12200.

- *loan_int_rate* - lãi suất



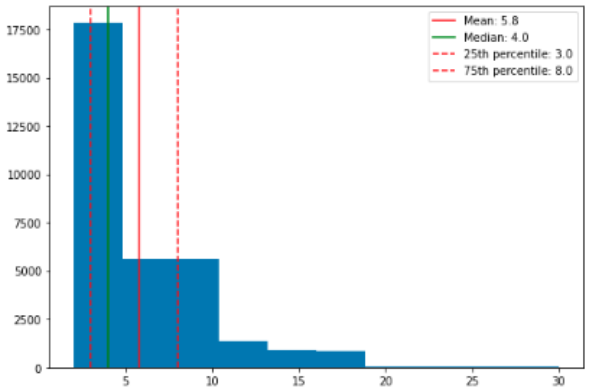
Nhận xét: Phân phối lãi suất lệch nhẹ phải và khá gần về phân phối chuẩn. Với giá trị trung bình gần trùng với giá trị trung vị. Lãi suất (%) tập trung trong khoảng 7.9 đến 13.47, và không xuất hiện nhiều điểm nhiễu.

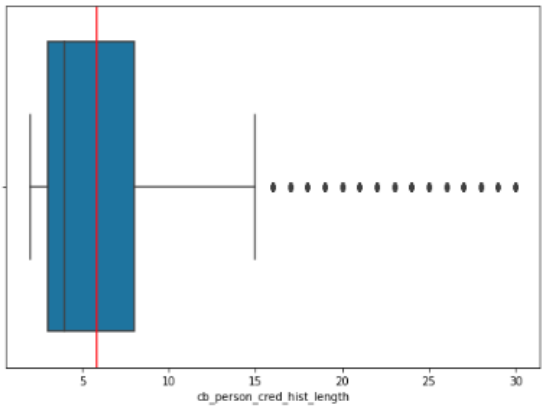
- *loan_percent_income* - tỷ lệ khoản vay vốn trên thu nhập



Nhận xét: Phân phối lãi suất lệch nhẹ phải và khá gần về phân phối chuẩn. Với giá trị tập trung trong khoảng 0.09 và 0.23.

- *cb_person_cred_hist_length* - số năm mở tín dụng



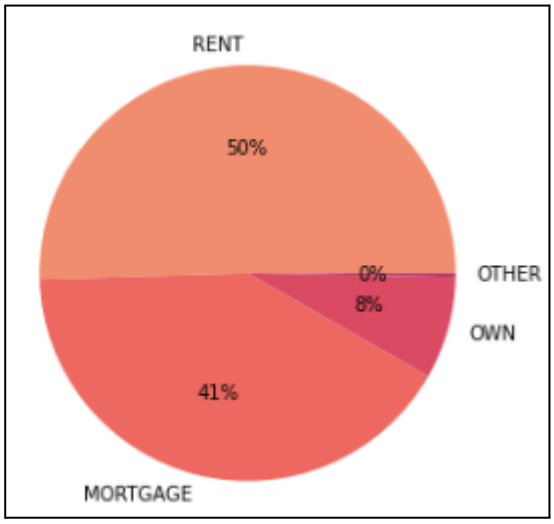


Nhận xét: Đồ thị lệch nhẹ phải, hầu hết tập trung trong khoảng từ 3 đến 8.

Các trường dạng hạng mục:

- *person_home_ownership* - phương thức sở hữu nhà ở

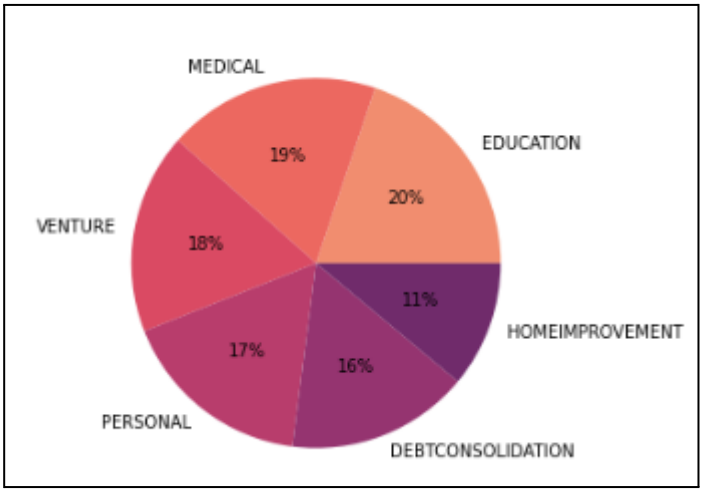
person_home_ownership	Frequencies	% Percentages
RENT	16446	50.48
MORTGAGE	13444	41.26
OWN	2584	7.93
OTHER	107	0.33



Nhận xét: tỷ lệ lớn nhất trong phương thức sở hữu nhà ở là thuê nhà (rent) và thế chấp tài sản (mortage), còn lại là chủ sở hữu và phương thức khác chiếm ít nhất.

- *loan_intent*- mục đích vay vốn

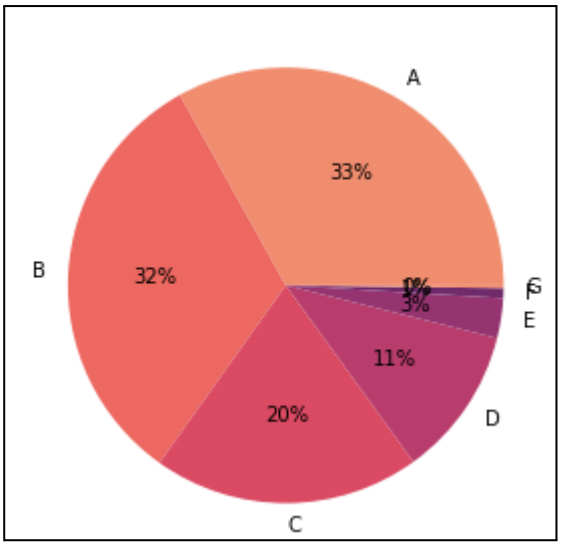
loan_intent	Frequencies	% Percentages
EDUCATION	6453	19.81
MEDICAL	6071	18.63
VENTURE	5719	17.55
PERSONAL	5521	16.95
DEBTCONSOLIDATION	5212	16.00
HOMEIMPROVEMENT	3605	11.06



Nhận xét: tỷ lệ mục đích vay vốn khá cân bằng giữa các hạng mục. Với lớn nhất là 20% cho mục đích giáo dục.

- *loan_grade*- xếp hạng vay vốn

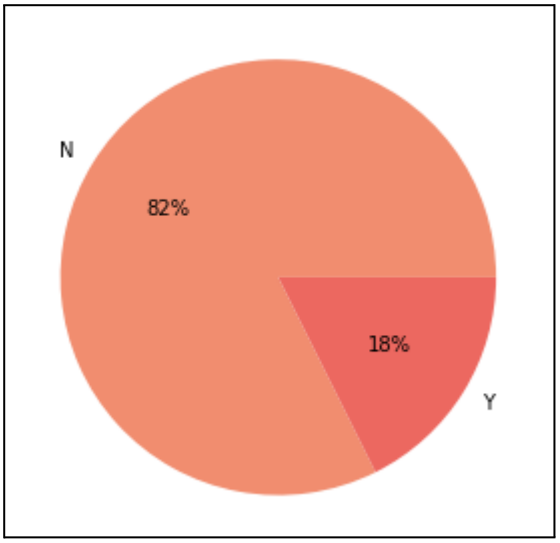
loan_grade	Frequencies	% Percentages
A	10777	33.08
B	10451	32.08
C	6458	19.82
D	3626	11.13
E	964	2.96
F	241	0.74
G	64	0.20



Nhận xét: Hầu hết xếp hạng cho vay của khách hàng tập trung trong các hạng mục A, B, C và D, số ít là các hạng mục E, F, G.

- *cb_person-default_on_file* - lịch sử vỡ nợ

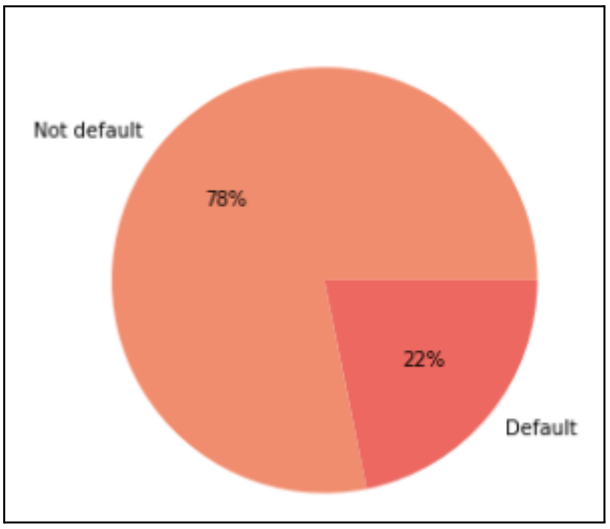
cb_person_default_on_file	Frequencies	% Percentages
N	26836	82.37
Y	5745	17.63



Nhận xét: Tỉ lệ khách hàng có lịch sử vỡ nợ thấp hơn khoảng 18%, còn lại là có lịch sử vay vốn trong sạch.

- *loan_status*- xếp hạng vay vốn (nhãn dự đoán)

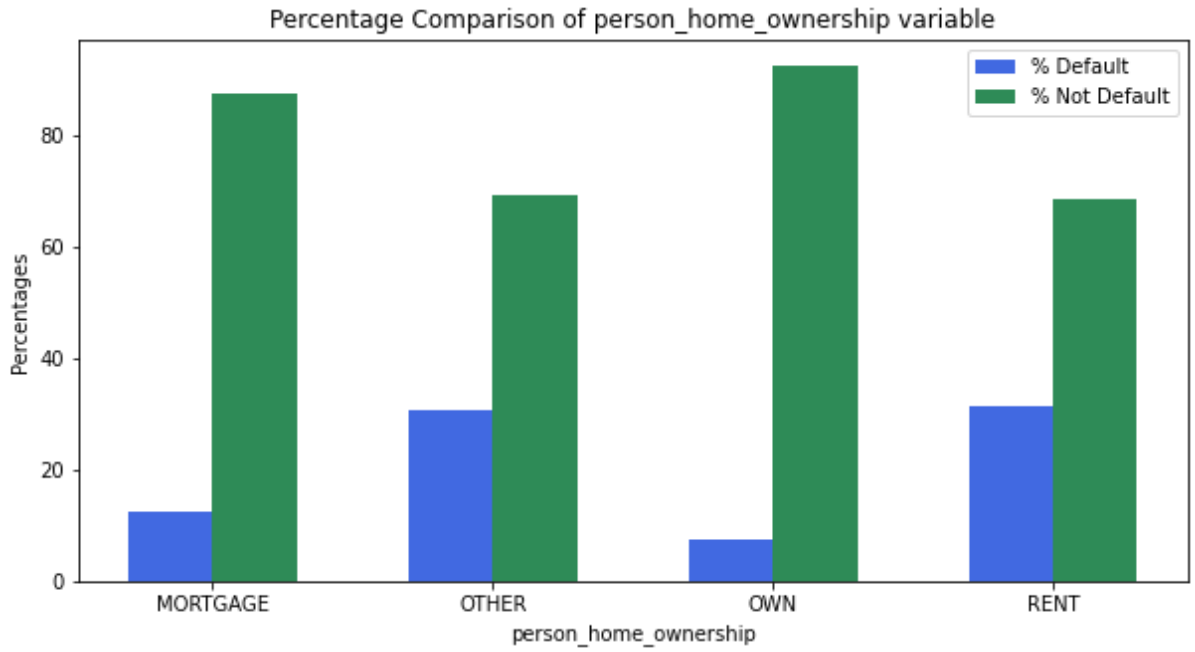
loan_status	Frequencies	% Percentages
Not default	25473	78.18
Default	7108	21.82



Nhận xét: Đây là nhãn dự đoán của bài toán. Với phần lớn khách hàng là không vỡ nợ (Not Default) 78%, 22% là vỡ nợ (Default).

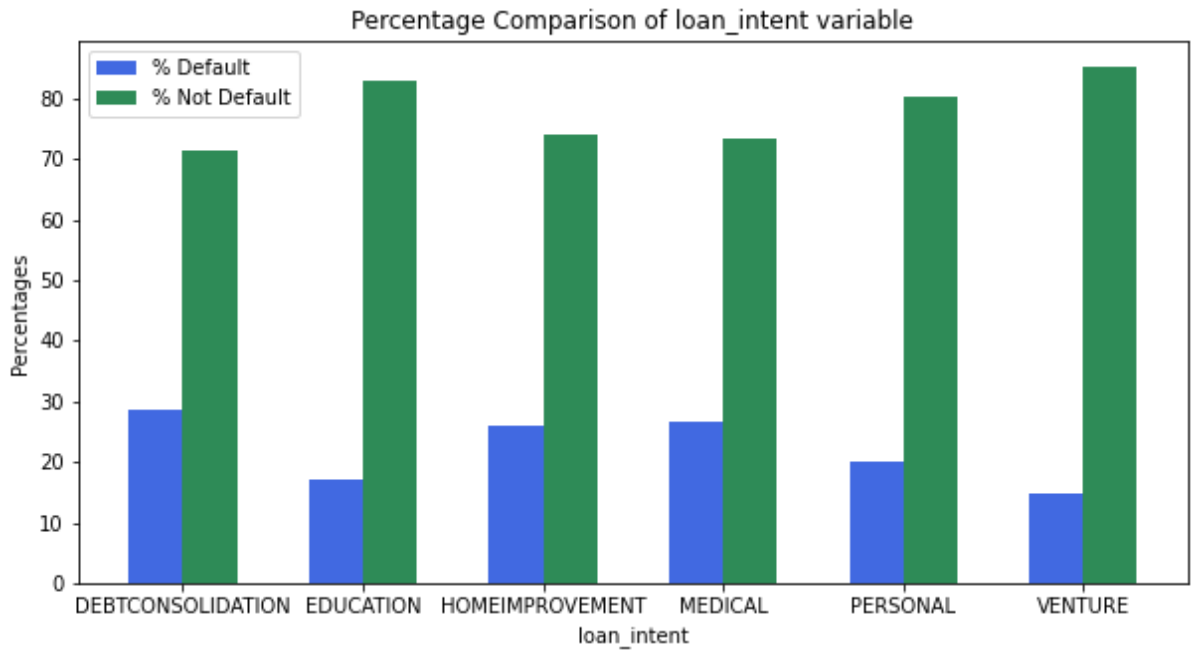
1.5. Mối tương quan giữa các trường dữ liệu dạng hạng mục với trường dự đoán.
1.5.1. Các trường dạng hạng mục

- Phương thức sở hữu nhà ở và tình trạng cho vay *(loan_status)*



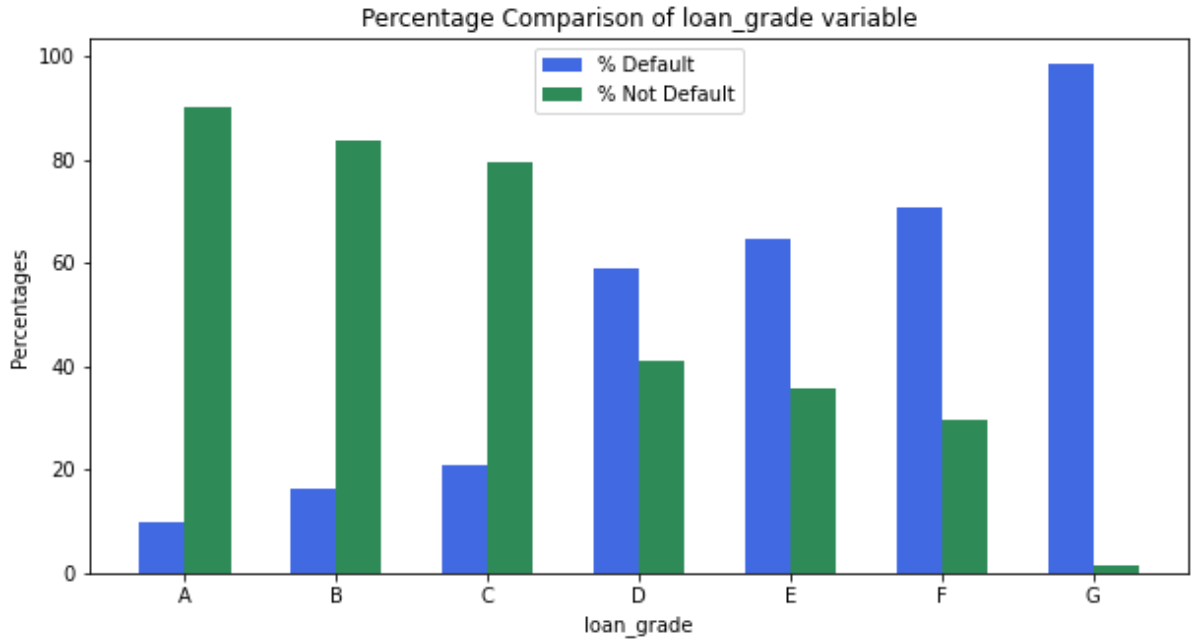
Nhận xét: Trong tổng số khách hàng vỡ nợ, phương thức ở hữu nhà ở bằng thế chấp (mortage) và chủ sở hữu (own) hiếm khi bị vỡ nợ, phương thức thuê nhà (rent) và các phương thức khác có xu hướng bị vỡ nợ nhiều hơn.

- Mục đích vay vốn và tình trạng cho vay *(loan_status)*



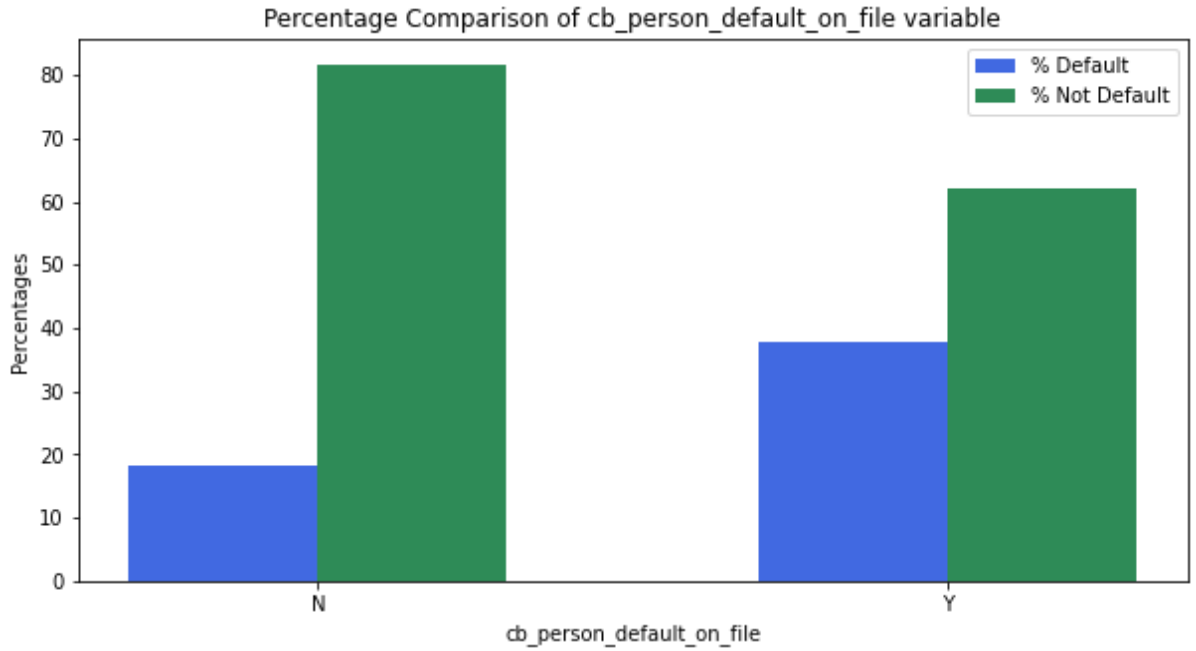
Nhận xét: Các khoản vay với mục đích lý do gộp nợ (DEBTCONSOLIDATION) và y tế (MEDICAL) bị vỡ nợ nhiều nhất. Khoản vay với mục đích giáo dục (EDUCATION) và liên doanh (VENTURE) khả năng vỡ nợ thấp nhất. Nhìn chung không có nhiều sự chênh lệch của các mục đích vay vốn trong tổng số khác hàng không bị vỡ nợ.

- Xếp hạng cho vay và tình trạng cho vay *(loan_status)*



Nhận xét: Khoản vay bị vỡ nợ tăng dần trong các danh mục xếp hạng cho vay từ A đến G. Khoản vay không bị vỡ nợ giảm dần từ A đến G. Điều này khá hợp lý vì các xếp hạng cho vay xếp hạng tăng dần mức độ không trả đc nợ của khách hàng. A, B, C, D, E… có thể là rất tốt, tốt, trung bình, nghi ngờ, thua lỗ, …

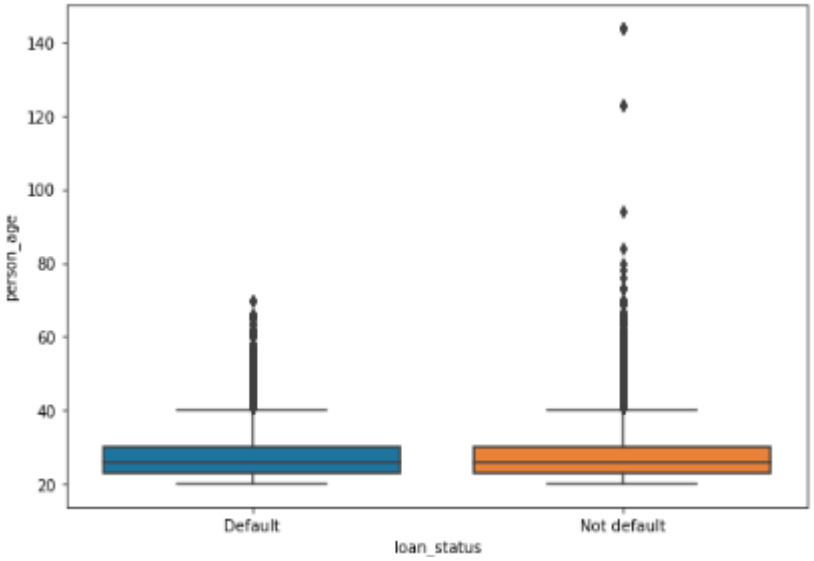
- lịch sử vỡ nợ và tình trạng cho vay *(loan_status)*



Nhận xét: Khách hàng có lịch sử vỡ nợ thường có tỉ lệ vỡ nợ cao hơn. Khách hàng có lịch sử trong sạch không vỡ nợ thường có khả năng trả được nợ cao hơn.

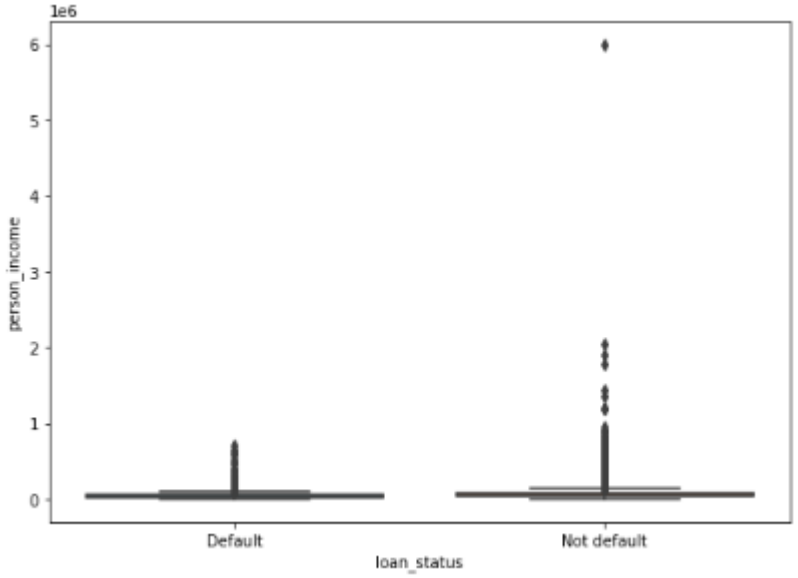
1.5.2. Các trường dạng số

- lịch sử vỡ nợ và tình trạng cho vay *(loan_status)*



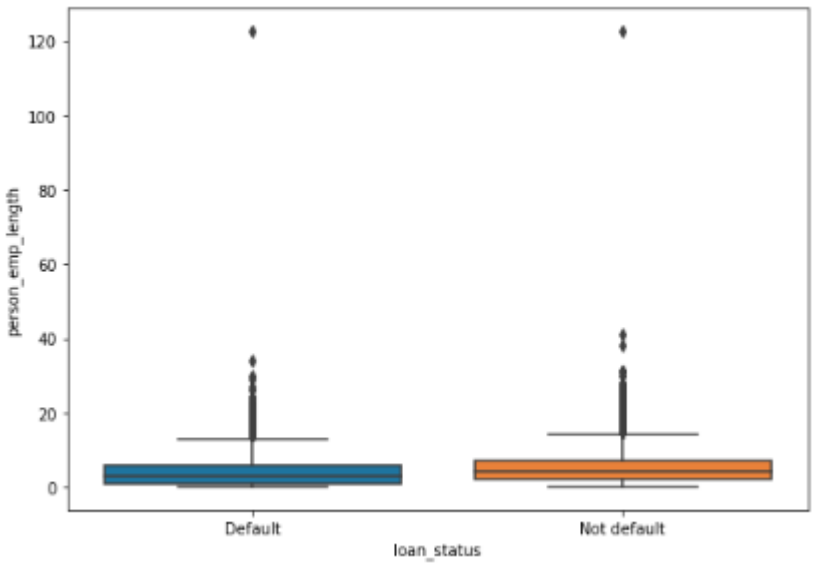
Nhận xét: Trong số khách hàng vỡ nợ và trả được nợ độ tuổi phổ biến trong khoảng 23 đến 30. Tuổi của khách hàng không giao động nhiều với khả năng vỡ nợ.

- Thu nhập hàng năm và tình trạng cho vay *(loan_status)*



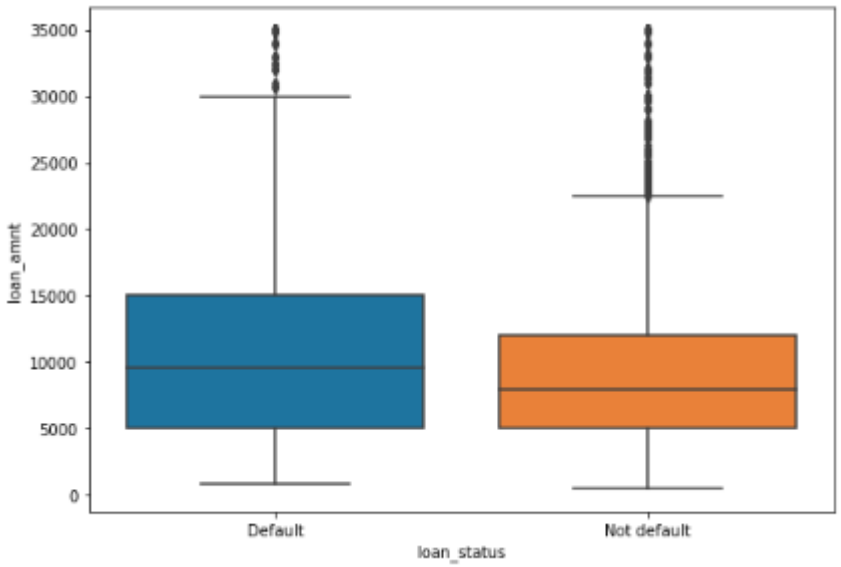
Nhận xét: Trong tổng số khách hàng vỡ nợ, khoảng thu nhập phổ biến từ 30,000 đến 59,497, với thu nhập trung bình khoảng 49,000. Trong tổng số khách hàng trả được nợ, thu nhập phổ biến từ 42,000 đến 84,000, với thu nhập trung bình khoảng 70,804. Khách hàng có thu nhập hàng năm cao thường có khả năng trả được nợ hơn so với khách hàng có thu nhập thấp hơn.

- số năm làm thuê và tình trạng cho vay *(loan_status)*



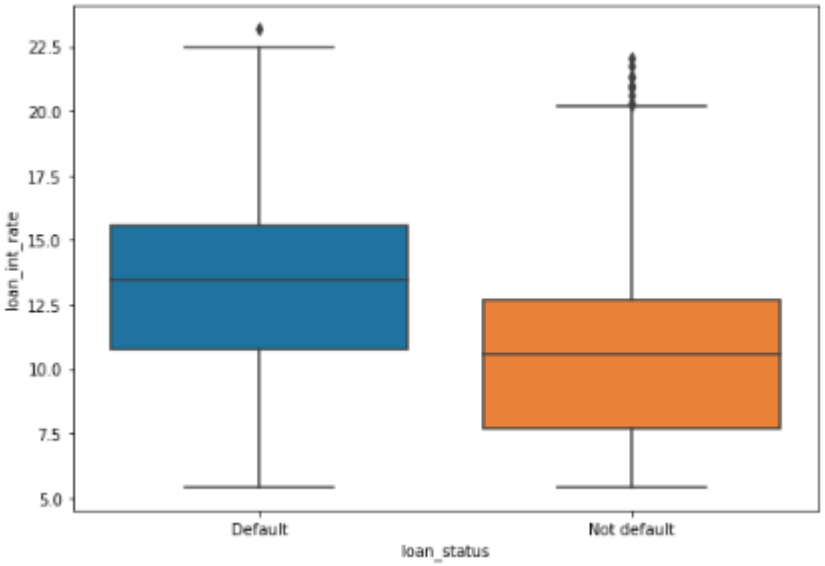
Nhận xét: Trong số khách hàng vỡ nợ, số năm làm thuê tập trung trong khoảng 1 đến 6 năm với số năm trung bình khoảng 4 năm. Trong số khách hàng trả được nợ, số năm làm thuê phần lớn trong khoảng 2 đến 7 năm.

- khoản vay và tình trạng cho vay *(loan_status)*



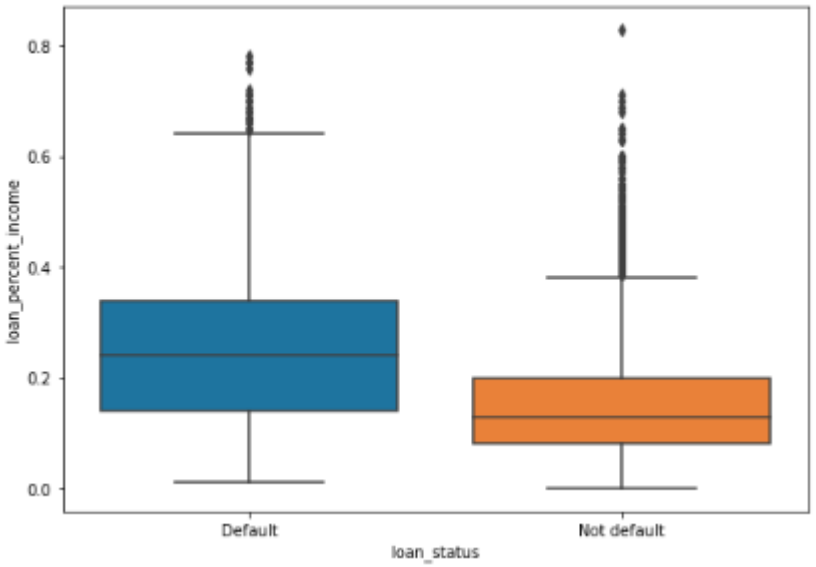
Nhận xét: Khách hàng có khoản vay cao có khả năng vỡ nợ cao hơn so với khoản vay thấp hơn.

- lãi suất và tình trạng cho vay *(loan_status)*



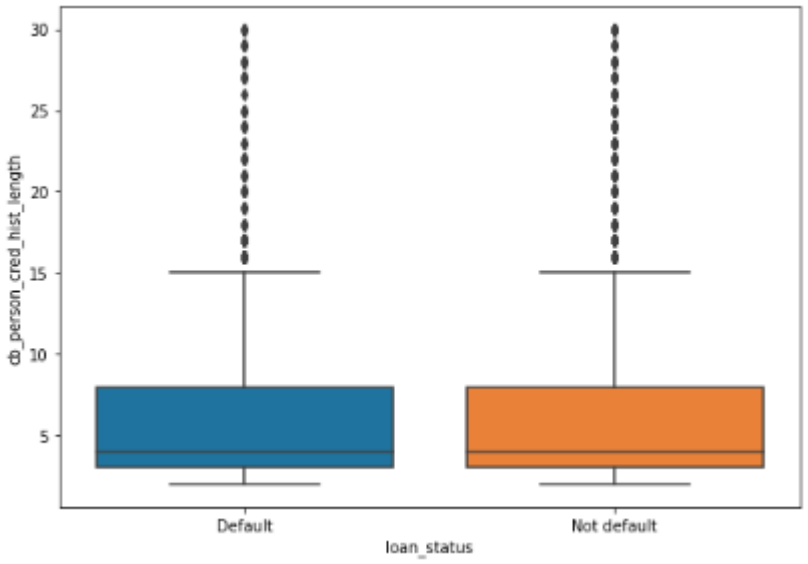
Nhận xét: Lãi suất vay vốn cao thì khả năng vỡ nợ cao hơn, lãi suất thấp.

- tỷ lệ khoản vay vốn trên thu nhập và tình trạng cho vay *(loan_status)*



Nhận xét: Tỷ lệ khoản vay trên thu nhập cao khả năng vỡ nợ của khách hàng cao hơn so với khả năng trả được nợ.

- số năm mở tín dụng và tình trạng cho vay *(loan_status)*

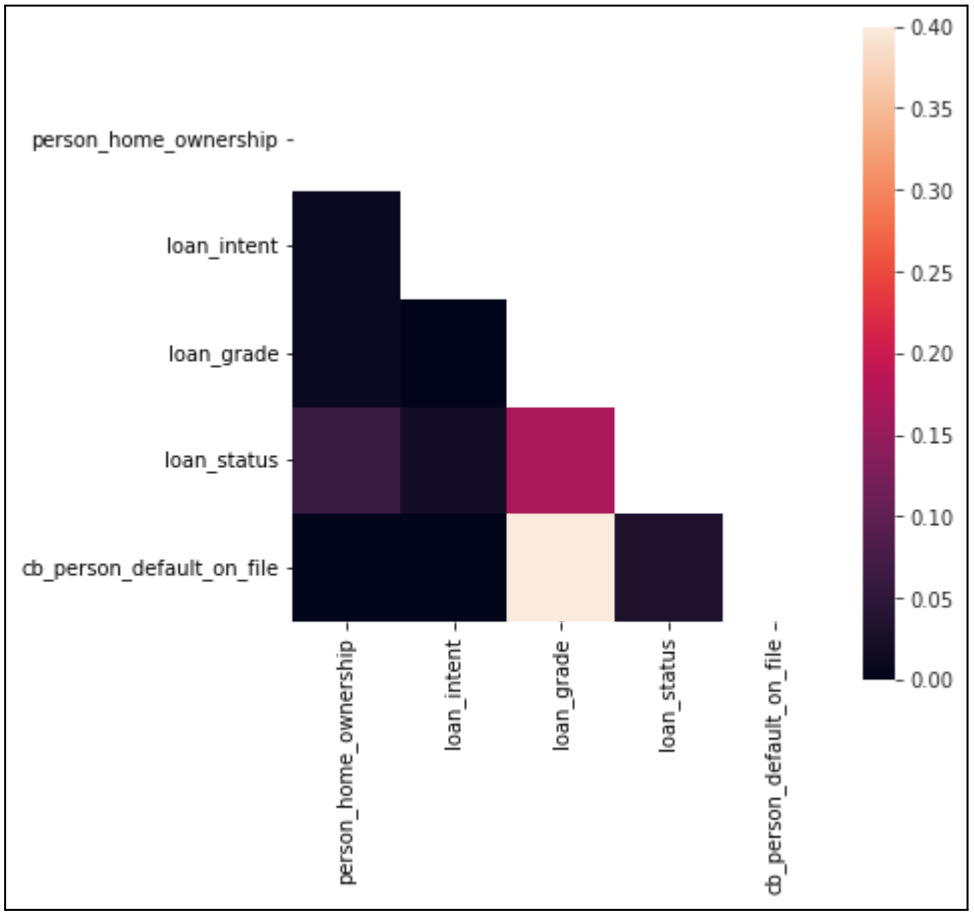


Nhận xét: Số năm mở tín dụng không làm dao động nhiều giữa khả năng vỡ nợ và trả được nợ.

1.6 Mối tương quan và liên hệ giữa các biến không phải cột nhãn

1.6.1 Mối tương quan giữa các biến dạng thư mục

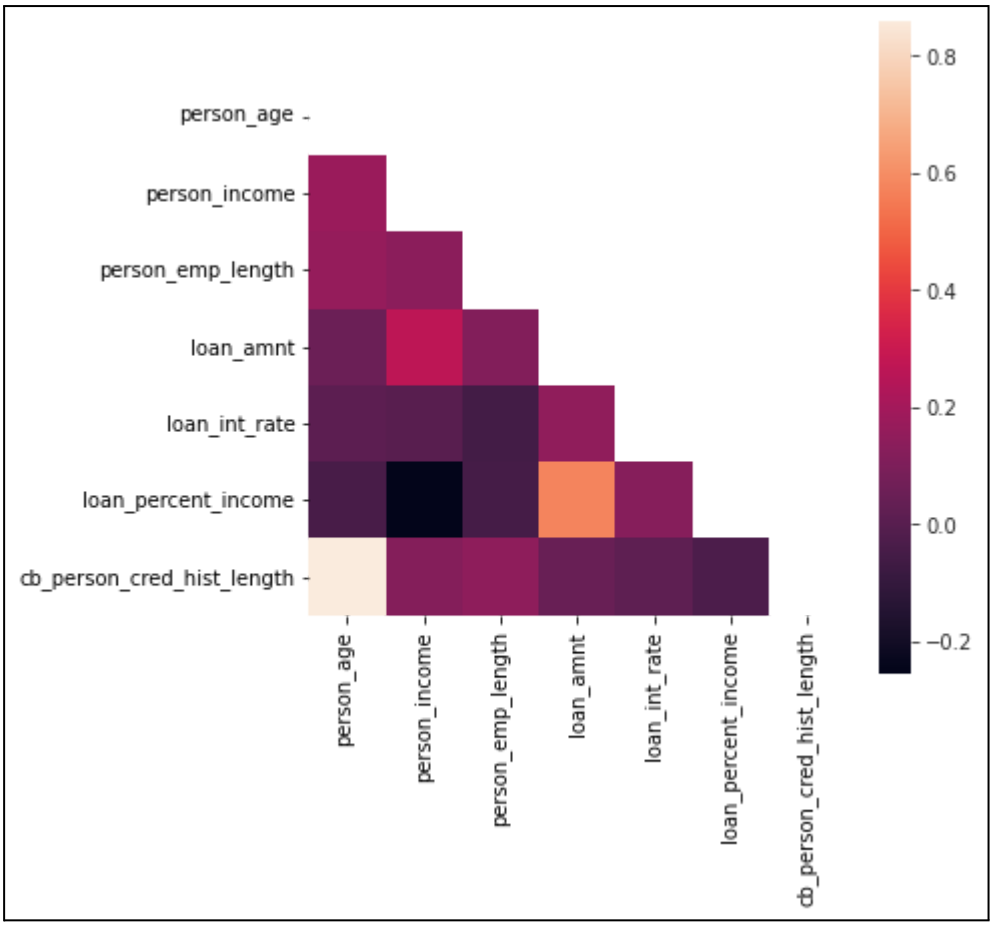
Sử dụng phép đo Cramer's V, có giá trị trong khoảng 0-1. Không có mối liên hệ hay tương quan giữa các biến khi giá trị này gần bằng 0, có sự tương quan mạnh khi giá trị này gần bằng 1.



Nhận xét: các cặp tình trạng vay vốn (loan_status) với phương thức sở hữu nhà ở (person_home_ownership), tình trạng vay vốn (loan_status) với xếp hạng vay vốn (loan_grade), xếp hạng vay vốn (loan_grade) với lịch sử vỡ nợ (cb_person_default_on_file) đều có sự tương quan nhẹ với nhau.

1.6.2 Mối tương quan giữa các biến dạng số

Sử dụng phép đo tương quan Pearson để đo mối liên hệ tuyến tính giữa hai biến dạng số. Có giá trị trong khoảng [-1, 1]. Không có mối tương quan khi giá trị gần bằng 0, có mối tương quan dương mạnh khi giá trị gần băng 1, có mối tương quan âm mạnh khi giá trị gần bằng -1.



Nhận xét:

- thu nhập hàng năm (person_income) có mối tương quan dương khá rõ rệt với tuổi của khách hàng. Tuổi càng lớn thì khả năng thu nhập càng cao.
- khoản vay (loan_amnt) có mỗi tương quan dương nhẹ với tỷ lệ khoản vay vốn trên thu nhập hàng năm (loan_percent_income) và thu nhập hàng năm (person_income).

2. Weight of Evidence - IV (trọng số giới thiệu - chỉ số giá trị thông tin)

Phương pháp xếp hạng các biến có liên quan đến biến nhãn dự đoán, độ quan trọng của biến độc lập với biến mục tiêu dựa vào chỉ số giá trị thông tin IV. Giá trị IV được tính toán thông qua trọng số giới thiệu WOE.
Giá trị IV:

- <= 0.02: Biến không có tác dụng trong việc phân loại hồ sơ Good/Bad
- 0.02 - 0.1: yếu
- 0.1 - 0.3: trung bình
- 0.3 - 0.5: mạnh
- => 0.5: Biến rất mạnh thể hiện mối quan hệ trực tiếp để định nghĩa hồ sơ good/bad.

col name	IV	rank
cb_person_cred_hist_length	0.004207	Useless
person_age	0.010719	Useless
person_emp_length	0.060641	Weak
loan_amnt	0.085864	Weak
loan_intent	0.095977	Weak
cb_person_default_on_file	0.164265	Medium
person_home_ownership	0.375582	Strong
person_income	0.469830	Strong
loan_int_rate	0.657280	suspicious
loan_percent_income	0.872763	suspicious
loan_grade	0.882659	suspicious

Hình: xếp hạng độ quan trọng của các biến đối với biến mục tiêu loan_status

Nhận xét:

- Biến *cb_person_cred_hist_length* và *person_age* không có tác dụng trong việc dự báo khả năng vỡ nợ của khách hàng.
- Các biến *person_emp_length*, *loan_amnt*, *loan_intent*, *cb_person_default_on_file* có tác dụng nhẹ.
- Các biến còn lại có tác động mạnh, liên quan trực tiếp đến việc dự đoán mục tiêu khả năng vỡ nợ.