ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH YẾU TỐ ẢNH HƯỞNG VÀ DỰ ĐOÁN KẾT QUẢ PHÊ DUYỆT KHOẢN VAY

Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Trần Đại Hiển	22520426	KHDL
2	Trần Lương Vân Nhi	22521044	KHDL

TP. HỒ CHÍ MINH – 06/2025

1. GIỚI THIỆU

Trong đồ án môn học này, chúng tôi đã áp dụng các kỹ thuật phân tích trực quan để khám phá dữ liệu, vận dụng các kiến thức để xây dựng mô hình học máy phù hợp để dự đoán thông tin chấp thuận khoản vay và trực quan dữ liệu để có thể rút ra các đánh giá cần thiết. Dựa trên bộ dữ liệu thông tin phê duyệt khoản vay lấy từ trên Kaggle, chúng tôi tiến hành tiền xử lý dữ liệu, sau đó sử dụng các thư viện sklearn, matplotlib.pyplot, seaborn tiến hành thực hiện phân tích thăm dò và để hiểu rõ các thuộc tính, các mối quan hệ giữa các đặc trưng và đánh giá những bất thường. Để có thể rút ra thêm các thông tin ngoài nhằm cho cái nhìn đa chiều với bài toán, chúng tôi trực quan dữ liệu bằng Power BI để có insight từ dữ liệu. Sau đó, chúng tôi sẽ thực hiện thử nghiệm trên 4 mô hình là Logistic Regression, K-Nearest Neighbor, Decision Tree, Random Forest, sử dụng độ đo Accuracy và Macro F1 Score để đánh giá bài toán. Kết quả cho ra mô hình phù hợp với bài toán và bộ dữ liệu nhất là Random Forest. Cuối cùng, chúng tôi xây dựng một framework, với mục đích thông qua API, người dùng có thể gửi các thông tin và trả về kết quả dự đoán.

Đề tài có sử dụng nội dung của dự án cùng tên, trong môn học Phân tích và Trực quan dữ liệu, mã môn học DS105 do thầy Phạm Thế Sơn hướng dẫn. Đồ án đạt 9 điểm trong học kỳ 1, năm học 2024-2025. Ngoài kết quả đã làm, đồ án này đã bổ sung thêm framework để có thể nhập dữ liệu và nhận kết quả dự đoán thông qua API.

2. MÔ TẢ DỮ LIỆU

Bộ dữ liệu kết hợp giữa thông tin rủi ro tín dụng và rủi ro tài chính nhằm cung cấp toàn vẹn dữ liệu về vay mượn và quyết định phê duyệt khoản vay của một công ty tài chính. Thông tin có thể khai thác bao gồm các thông tin cá nhân liên quan đến nhóm tuổi, trình độ học vấn, công việc của người vay và các thông tin liên quan đến khoản vay như số tiền, lãi suất, thời gian vay mượn,… và tình trạng chấp thuận/từ chối.

Bộ dữ liệu gốc bao gồm 14 cột (thuộc tính) và 45000 dòng, trong đó có 9 thuộc tính là biến số và 5 thuộc tính là biến phân loại. Biến phụ thuộc ‘status’ cho biết kết quả phê duyệt khoản vay của người đó.

Bộ dữ liệu được thi thập bởi tác giả Tawei Lo.

Mô tả chi tiết các cột thuộc tính:

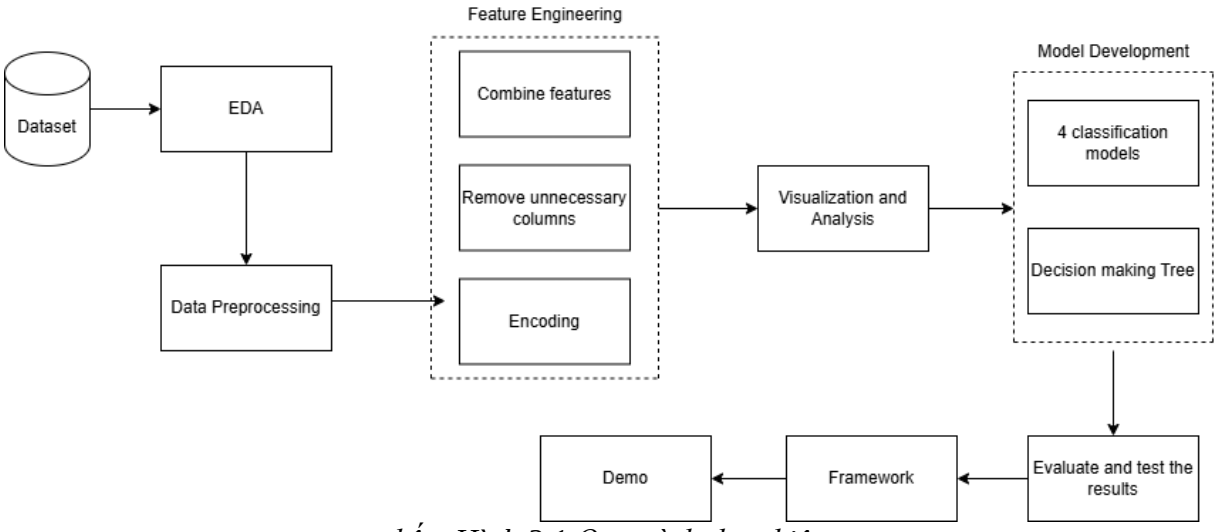
STT	Tên thuộc tính	Kiểu dữ liệu	Mô tả
1	person_age	float	Tuổi của người vay
2	person_gender	string	Giới tính của người vay
3	person_education	string	Trình độ học vấn của người vay
4	person_income	float	Thu nhập hằng năm của người vay

5	person_emp_exp	integer	Số năm kinh nghiệm trong công việc của người vay
6	person_home_ownership	string	Tình trạng sở hữu nhà của người vay
7	loan_amnt	float	Số tiền muốn vay
8	loan_intent	string	Mục đích của khoản vay
9	loan_int_rate	float	Lãi suất vay mượn
10	loan_percent_income	float	Số tiền vay dưới dạng phần trăm thu nhập hằn năm
11	cb_person_cred_hist_length	float	Lịch sử tín dụng (theo năm)
12	credit_score	integer	Điểm tín dụng của người vay
13	previous_loan_defaults_on_file	string	Chỉ số về các khoản vay trước đó
14	loan_status	integer	Trạng thái phê duyệt (1: chấp nhận, 0: từ chối)

Hình 2.1: Mô tả chi tiết thuộc tính

3. QUY TRÌNH THỰC HIỆN

Chúng tôi thực hiện phân tích dữ liệu theo quy trình như sau:



kết qHình 3.1 Quy trình thực hiện

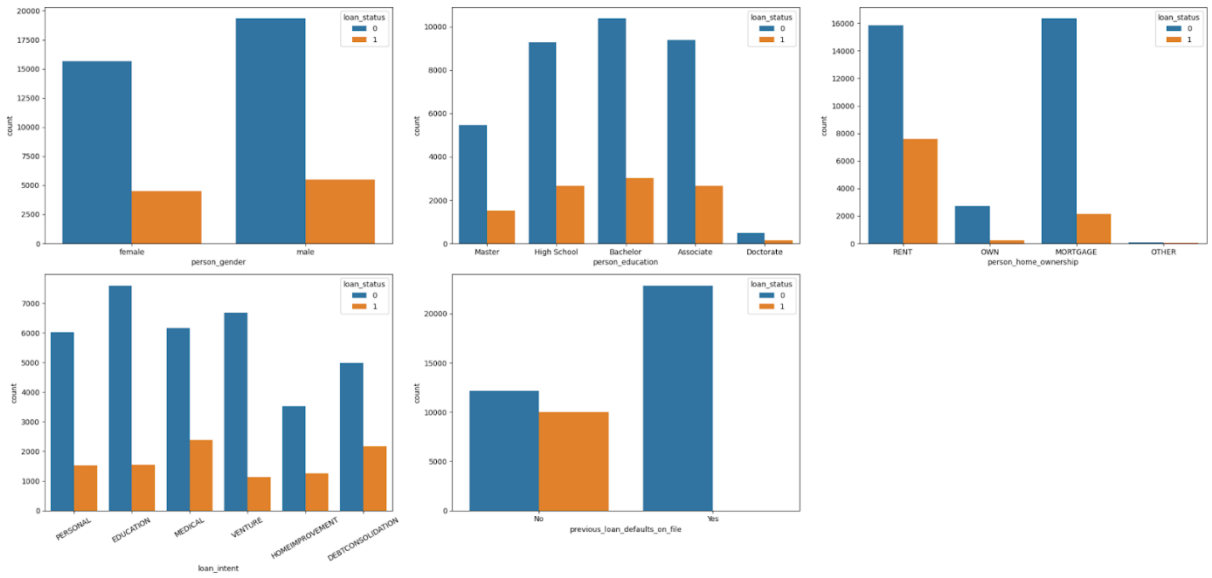
4. NỘI DUNG PHÂN TÍCH

4.1. Thăm dò sơ bộ

4.1.1. Biến phân loại

Hình 4.1 Trực quan hóa các biến phân loại

Countplox cho thấy các thuộc tính person_gender, loan_intent, previous_loan_defaults_on_file khá cân bằng, còn các biến person_education, persion_home_ownership lại bị mất cân bằng ở một số cột.

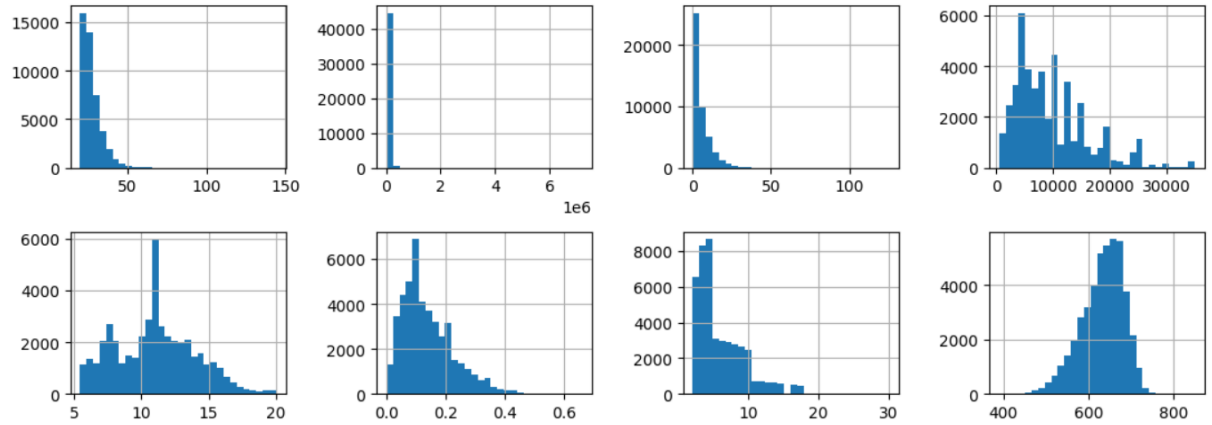


Hình 4.2 Trực quan hóa các biến phân loại

Chỉ có một chút khác biệt không đáng kể giữa hai giới tính trong việc phê duyệt khoản vay, số hồ sơ được duyệt có vẻ đồng đều so với trình độ học vấn của người vay, và có sự chênh lệch giữa các mục tiêu cho vay nhưng không đáng kể, nên các thuộc tính này có vẻ không phải yếu tố mạnh mẽ ảnh hưởng đến kết quả cho vay.

Trong khi đó, người có tình trạng nhà đang thế chấp dường như có tỷ lệ bị từ chối cao hơn so với những nhóm khác. Nó có thể gợi ý rằng họ có thể có mức độ ổn định tài chính kém hơn chủ sở hữu nhà. Những người từng có lịch sử vỡ nợ bị từ chối cao hơn hẳn nhóm còn lại. Đặc điểm này có ảnh hưởng mạnh mẽ đến việc đồng ý cho vay, vì khoản nợ xấu trong quá khứ báo hiệu rủi ro tương tự cho lần xem xét hồ sơ hiện tại. Đây là những thuộc tính cần lưu ý lựa chọn khi xây dựng mô hình.

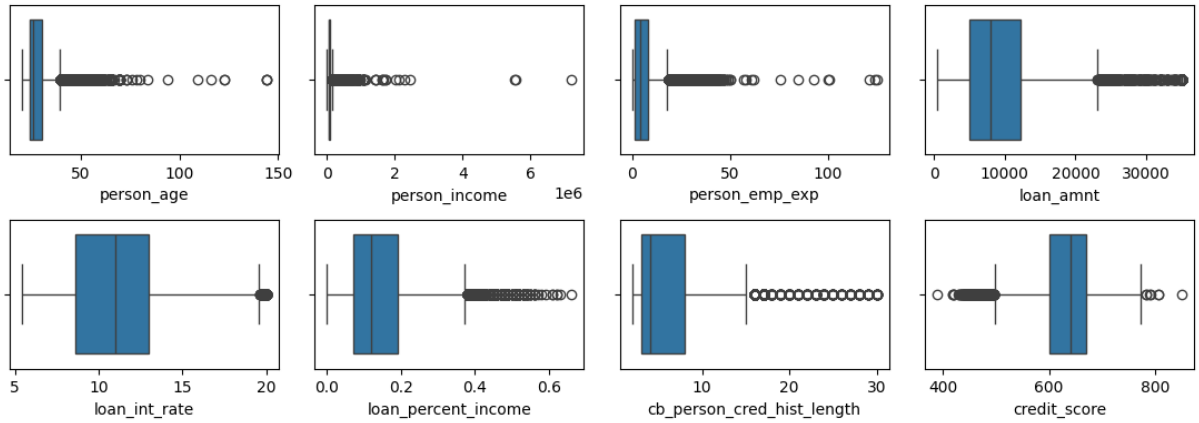
4.1.2 Biến liên tục



Hình 4.3 Trực quan hóa các biến liên tục

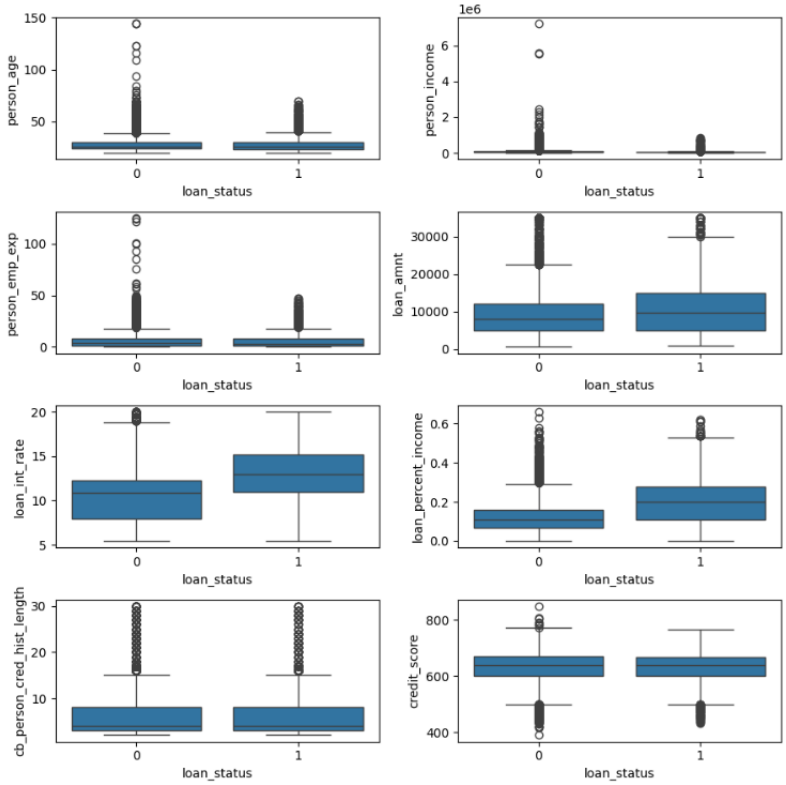
Các biến số của dữ liệu đa số lệch phải từ khá đến lệch mạnh. Số tiền vay có biểu đồ lệch phải mạnh, nên hầu hết người vay yêu cầu các khoản vay nhỏ và chiếm tỷ lệ khoảng dưới 20% thu nhập.

Riêng điểm tín dụng được phân phối khá chuẩn ở mức trung bình (600-700).



Hình 4.4 Trực quan hóa các biến liên tục

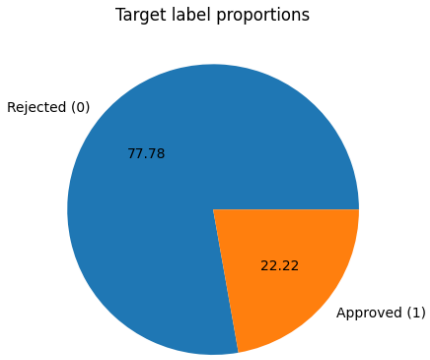
Boxplot cho thấy: Đa số hồ sơ vay tiền đến từ người trẻ (20-30 tuổi), nhưng có một số outliers lớn hơn 100 tuổi là một sự bất thường trong dữ liệu. Số năm kinh nghiệm làm việc cũng cho thấy nhiều người có số năm thấp dưới 10 năm, nhưng có nhiều outliers với giá trị rất lớn. Phần trăm thu nhập được sử dụng để trả khoản vay tập trung ở mức dưới 20%.



Hình 4.5 Trực quan hóa các biến liên tục

Nhưng các biến person_age, person_emp_exp, cb_person_cred_hist_length, credit_score này chồng lấn lên nhau nên có thể sẽ không có ảnh hưởng đáng kể đến biến mục tiêu (loan_status).

4.1.3 Biến mục tiêu



Hình 4.6 Trực quan hóa biến mục tiêu

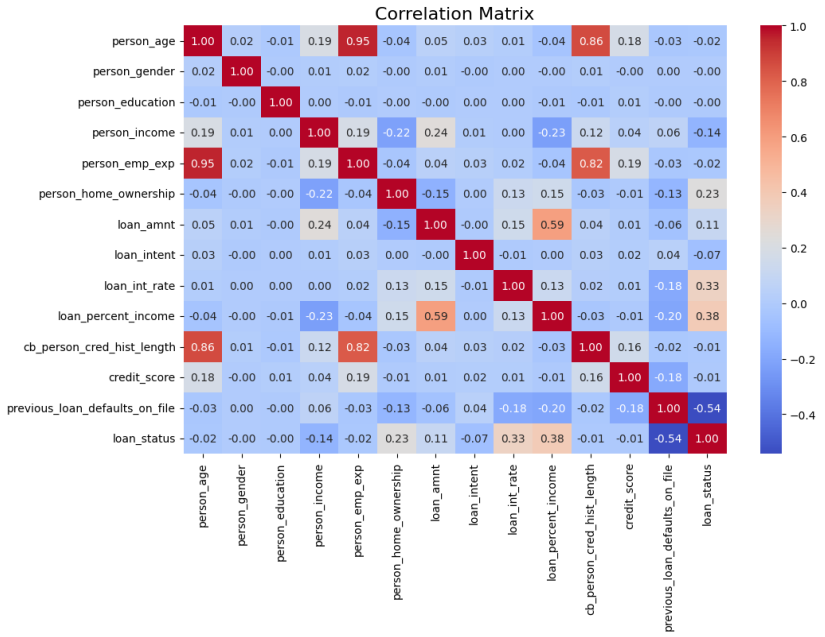
Dùng pie chart, dễ thấy có sự mất cân bằng dữ liệu khi đa số hồ sơ vay bị từ chối.

4.2. Thay thế các giá trị ngoại lai không phù hợp

Sau khi thay thế outliers hơn 100 tuổi của thuộc tính person_age với giá trị median(), tuổi lớn nhất hiện là 94, là một khoảng hợp lý cho tuổi người, duy trì median vẫn rơi vào 26 tuổi.

4.3. Feature Engineerings

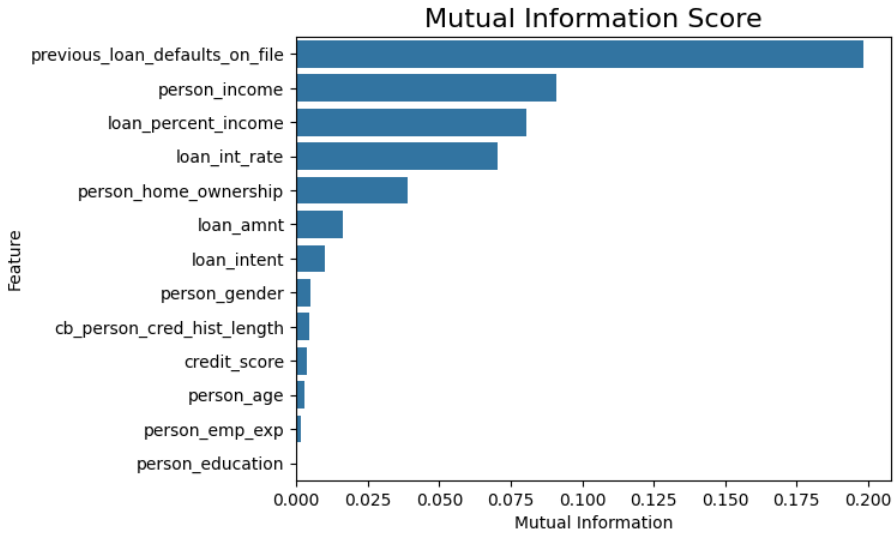
4.3.1. Kết hợp các nhóm có tương quan cao



Hình 4.7 Heatmap

Biến person_emp_exp có tương quan mạnh với biến person_age, nên kết hợp 2 cột thuộc tính bằng cách chọn person_age do đã xử lý outliers của thuộc tính này. 2

thuộc tính person_gender và person_education không có tương quan với biến mục tiêu nên bỏ những biến này đi.



Hình 4.8 Biểu đồ MIS

Dựa vào biểu đồ MIS, previous_loan_defaults_on_file là đặc trưng quan trọng nhất (~0.2) để dự đoán biến mục tiêu. Điều này hợp lý vì việc từng có nợ xấu thường ảnh hưởng lớn đến quyết định cho vay trong tương lai. Thu nhập cá nhân cũng là một đặc trưng quan trọng thứ hai, thể hiện mối quan hệ giữa khả năng tài chính và khả năng trả nợ. loan_percent_income và loan_int_rate: Cả hai thuộc tính này cũng có mức độ MIS tương đối, cho thấy vai trò của mức độ gánh nặng tài chính (khoản vay so với thu nhập) và lãi suất vay trong việc dự đoán kết quả. Và cũng cần quan tâm đến tình trạng sở hữu nhà của người đi vay.

Các đặc trưng cb_person_cred_hist_length, credit_score, person_age, person_emp_exp cho thấy một sự tương quan quá yếu với biến mục tiêu nên sẽ loại bỏ để đơn giản cho mô hình.

Các hiểu biết này hoàn toàn hợp lý với những gì đã phân tích sơ bộ ở trên.

4.3.2 Mã hóa dữ liệu

Chúng tôi sử dụng Label Encoder cho biến "previous_loan_defaults_on_file" vì biến chỉ có 2 giá trị "Yes" và "No", tương ứng với 1 và 0, nên không cần dùng One Hot Encoder. Từ đó sẽ tiết kiệm bộ nhớ và làm giảm kích thước của bộ dữ liệu.

Đối với các biến phân loại còn lại ("person_home_ownership", "loan_intent"), sẽ dùng cách mã hóa One Hot. Nguyên nhân là vì 2 biến này có nhiều hơn 2 giá trị ("person_home_ownership" - 4, "loan_intent" - 6), nếu sử dụng cách mã hóa Label thì sẽ vô tình tạo nên mối quan hệ thứ bậc cho 2 biến này (4 > 3 > 2 > 1).

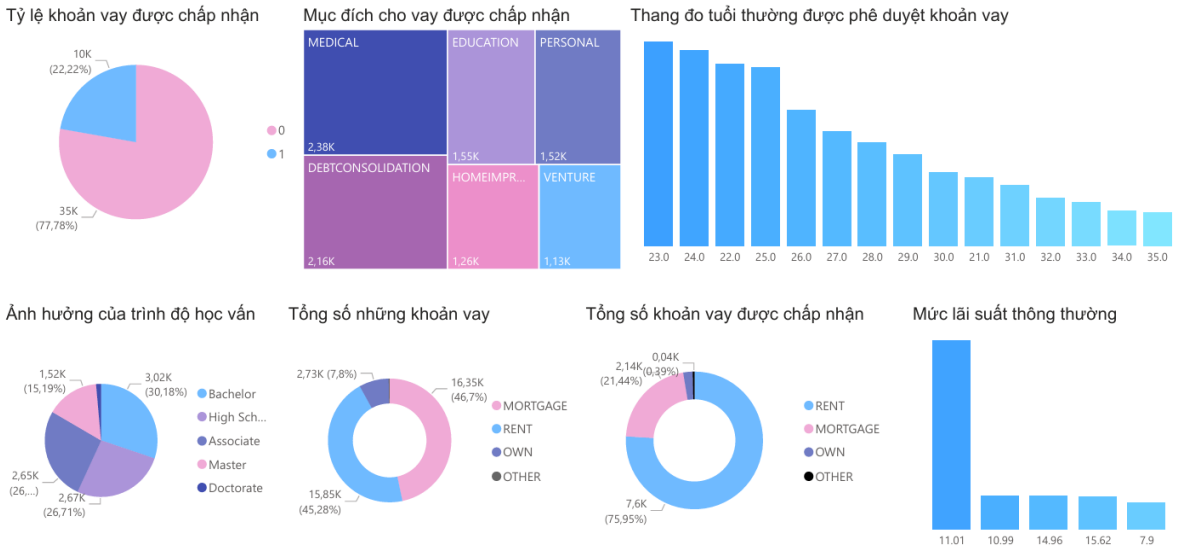
4.3.3 Standard Scaler

Để mô hình hội tụ nhanh hơn nên nhóm chúng tôi quyết định sử dụng StandardScaler để scale những biến liên tục lại. Đồng thời, chúng tôi sao chép dữ liệu

trước khi được scale, để khi xem xét mô hình Decision Tree đưa ra quyết định như thế nào, chúng tôi có thể tiện theo dõi giá trị thực của các biến liên tục trực quan hơn.

5. PHÂN TÍCH NỘI DUNG DỮ LIỆU

Ngoài ra để biết thêm thông tin về bộ dữ liệu, có thể thực hiện trực quan hóa bằng Power BI và nhận thấy được những thông tin nổi bật như:



Hình 5.1 Trực quan dữ liệu bằng Power BI

- Tỷ lệ khoản vay được chấp nhận: 77.78% khoản vay bị từ chối (0), chỉ 22.22% trong tổng số được chấp thuận. Điều này cho thấy tỷ lệ xét duyệt và chấp thuận khoản vay của công ty tài chính này khá gắt.

- Các khoản vay với mục đích y tế sẽ được chấp thuận nhiều nhất, khoản vay với mục đích gộp nợ các khoản trong vay mượn cũng là một trong lý do được chấp thuận, tiếp theo là các khoản liên quan đến giáo dục, cá nhân, nhà cửa, mục đích vay để đầu tư với mức rủi ro sẽ ít được chấp thuận nhất.

- Khoảng tuổi 23-27 chiếm tỷ lệ cao nhất trong việc được phê duyệt khoản vay, và các độ tuổi cao hơn sẽ giảm dần tỉ lệ. Xu hướng này có thể chỉ ra nhóm tuổi trẻ dễ tiếp cận khoản vay hơn.

- Thường những người có bằng cử nhân có tỷ lệ được chấp thuận các khoản vay cao nhất, tiếp theo là cấp ba và hệ liên thông. Học vị tiến sĩ và thạc sĩ chiếm tỷ trọng ít nhất. Có thể thấy trình độ học vấn cao không hoàn toàn quyết định khả năng vay được chấp nhận, để xét về vấn đề này cần xét thêm các biến tương quan với nó.

- Trong tổng số các khoản vay, có thể thấy người đang thế chấp nhà và người thuê nhà đều chiếm một phần khá đông. Nhưng xét đến những khoản vay được chấp thuận thì có thể thấy tỷ lệ được chấp thuận ở những người đang thuê nhà cao hơn hẳn những người đang thế chấp nhà. Đây là một trong những điểm rất nổi bật mà khi xét nội dung cần lưu ý.

- Mức lãi suất thông thường được phê duyệt là 11.01, một mức lãi suất trung bình cao nhưng rất phổ biến so với các mức còn lại

6. PHÁT TRIỂN MÔ HÌNH VÀ ĐÁNH GIÁ KẾT QUẢ

6.1. Lựa chọn mô hình và độ đo đánh giá

- 4 mô hình phân loại: Linear Regression, KNN, Decision Tree và Random Forest.

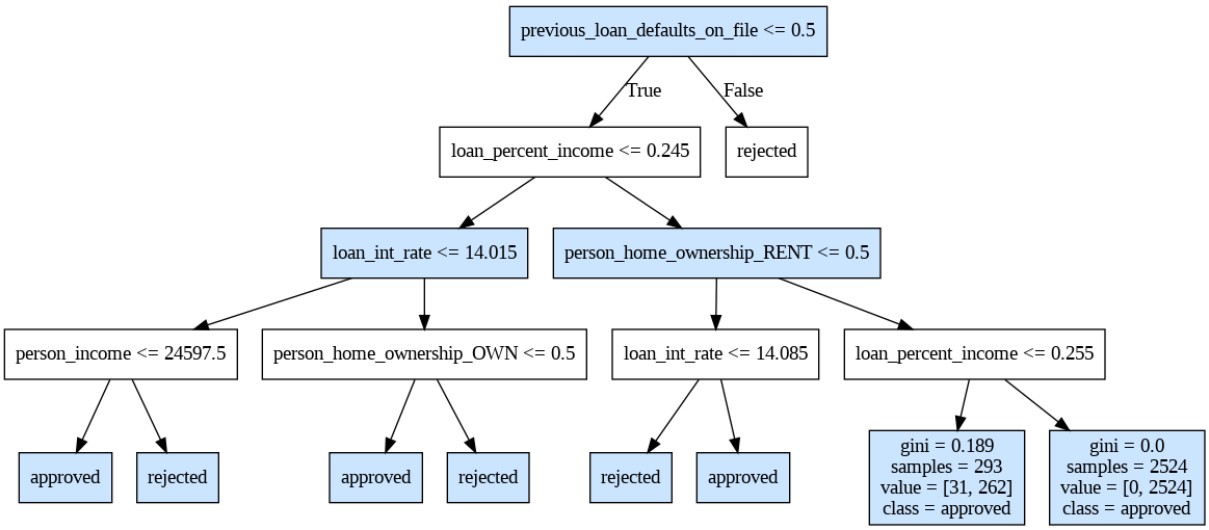
- Độ đo đánh giá: macro F1-score và accuracy. Do bài toán có sự mất cân bằng giữa 2 lớp dữ liệu nên chúng tôi tập trung xếp hạng mô hình dựa vào F1-score.

6.2. Kết quả và đánh giá

Mô hình	Macro F1-score	Accuracy
Logistic Regression	0.84	0.89
KNN	0.87	0.91
Decision Tree	0.87	0.91
Random Forest	0.90	0.93

Hình 6.1 Kết quả thực nghiệm

Với mô hình Decision Tree, degree được chọn là 4 để vừa đạt được độ chính xác cao, vừa đủ đơn giản để giải thích quyết định đồng ý hoặc từ chối hồ sơ vay tiền, hữu ích cho bên cho vay là người sử dụng mô hình, hiểu sâu sắc hơn về quyết định cho vay của mình.



Hình 6.2 Mô hình Decision Tree

Cây trên thể hiện lịch sử nợ xấu là yếu tố hàng đầu để có xem xét tiếp hồ sơ không. Nếu đã từng có nợ xấu thì khả năng cao hồ sơ nên được từ chối. Nếu chưa có lịch sử nợ xấu, mức vay hợp lý với thu nhập thì trong các trường hợp sau có thể tiếp nhận hồ sơ vay:

- Lãi suất vay ở mức phổ biến, và thu nhập thấp.

- Lãi suất vay cao hơn bình thường và người vay hiện không sở hữu nhà. Điều này có vẻ không hợp lý vì người đã có nhà đối mặt với ít rủi ro chi trả hơn. Chúng ta sẽ cần dữ liệu nhân khẩu học chuyên sâu hơn nếu cần hiểu hành vi của nhóm người này.
- Người vay hiện không thuê nhà và lãi suất cho vay cao hơn mức phổ biến. Điều này có thể gợi ý rằng nhóm người không thuê nhà có thể chấp nhận mức rủi ro tín dụng cao hơn.
- Tiếp nhận hồ sơ của người vay hiện đang thuê nhà, nhưng khả năng chấp nhận sẽ cao hơn với hồ sơ yêu cầu mức vay từ 0.245-0.255.

Các yếu tố đưa ra quyết định này phù hợp với những nhận định khi chúng tôi xem xét countplox của các biến previous_loan_defaults_on_file và persion_home_ownership và MIS của các biến loan_percent_income và loan_int_rate

6.3. Cross-Validation

Chúng tôi tiếp tục sử dụng kiểm chứng khả năng tổng quát của mô hình khi dùng K-Fold Cross-Validation. Kết quả có được sau khi thực hiện như sau:

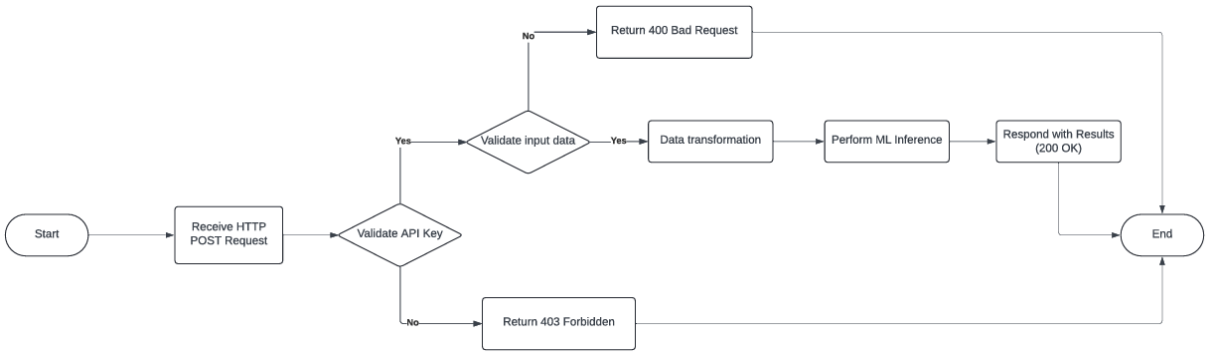
Mô hình	Macro F1-score	Accuracy
Logistic Regression	0.85	0.90
KNN	0.87	0.91
Decision Tree	0.87	0.91
Random Forest	0.90	0.93

Hình 6.3 Kết quả thực nghiệm với Cross-Validation

7. XÂY DỰNG FRAMEWORK

Chúng tôi thực hiện xây dựng một API RESTful dựa trên Django, sử dụng các mô hình học máy đã thực nghiệm ở phần trước để dự đoán khả năng phê duyệt khoản vay dựa trên thông tin người vay. Dữ liệu đầu vào được xử lý qua một pipeline bao gồm xác thực, chuyển đổi, dự đoán và trả kết quả.

7.1 Luồng xử lý



Hình 7.1 Sơ đồ luồng xử lý

Bước 1: Tiếp nhận yêu cầu (Request Reception)

- Client gửi yêu cầu HTTP POST chứa dữ liệu đăng ký khoản vay
- Hệ thống nhận yêu cầu từ client và phân tích payload dưới dạng JSON
- Chuyển đến Bước 2

Bước 2: Xác thực API Key (API Key Validation)

- Sử dụng decorator tùy chỉnh để kiểm tra tính hợp lệ của API key trong header của yêu cầu
- **Nếu API key không hợp lệ:** Hệ thống trả về lỗi 403 Forbidden và kết thúc luồng xử lý
- **Nếu API key hợp lệ:** Chuyển đến Bước 3

Bước 3: Xác thực dữ liệu đầu vào (Input Data Validation)

- Kiểm tra dữ liệu đầu vào để đảm bảo tất cả các trường bắt buộc đều có mặt trong payload
- Kiểm tra kiểu dữ liệu để xác nhận các trường số/chuỗi hợp lệ
- **Nếu dữ liệu không hợp lệ:** Hệ thống trả về lỗi 400 Bad Request và kết thúc luồng xử lý
- **Nếu dữ liệu hợp lệ:** Chuyển đến Bước 4

Bước 4: Tiền xử lý dữ liệu (Data Transformation)

- Kiểm tra các trường có dữ liệu dạng phân loại (categorical) và tự động chỉnh sửa các giá trị không hợp lệ về chuẩn
- Chuyển đổi đặc trưng bằng cách sử dụng one-hot encoding
- Chuẩn hóa giá trị số bằng StandardScaler đã được huấn luyện trước để chuẩn hóa các đặc trưng dạng số
- Chuyển đến Bước 5

Bước 5: Suy luận mô hình (ML Inference)

- Tải mô hình đã chọn (hoặc mặc định KNN) từ file pickle đã lưu trữ sẵn
- Dữ liệu đầu vào sau khi tiền xử lý được đưa vào mô hình để thực hiện suy luận
- Mô hình trả về kết quả dự đoán dạng nhị phân với 0 là "Từ chối khoản vay" và 1 là "Chấp thuận khoản vay"
- Chuyển đến Bước 6

Bước 6: Xử lý phản hồi (Response Processing)

- Chuyển kết quả dự đoán dạng số thành mô tả dễ hiểu
- Phản hồi JSON với mã **200 OK** trả về client bao gồm giá trị (value) và mô tả kết quả dự đoán (description)
- Kết thúc luồng xử lý

7.2 Input và Output

Client gửi một yêu cầu gồm các thông tin theo chuẩn sau:

Field	**Type**	**Range/Options**	**Description**
person_income	Number	12,000 - 72,000	Annual income of applicant
person_home_ownership	String	RENT, OWN, MORTGAGE, OTHER	Home ownership status
loan_amnt	Number	1,000 - 35,000	Requested loan amount
loan_intent	String	PERSONAL, EDUCATION, MEDICAL, VENTURE, HOMEIMPROVEMENT, DEBTCONSOLIDATION	Purpose of loan
loan_int_rate	Number	10.0 - 20.0	Annual interest rate (%)
loan_percent_income	Number	0.01 - 0.66	Loan amount as percentage of income

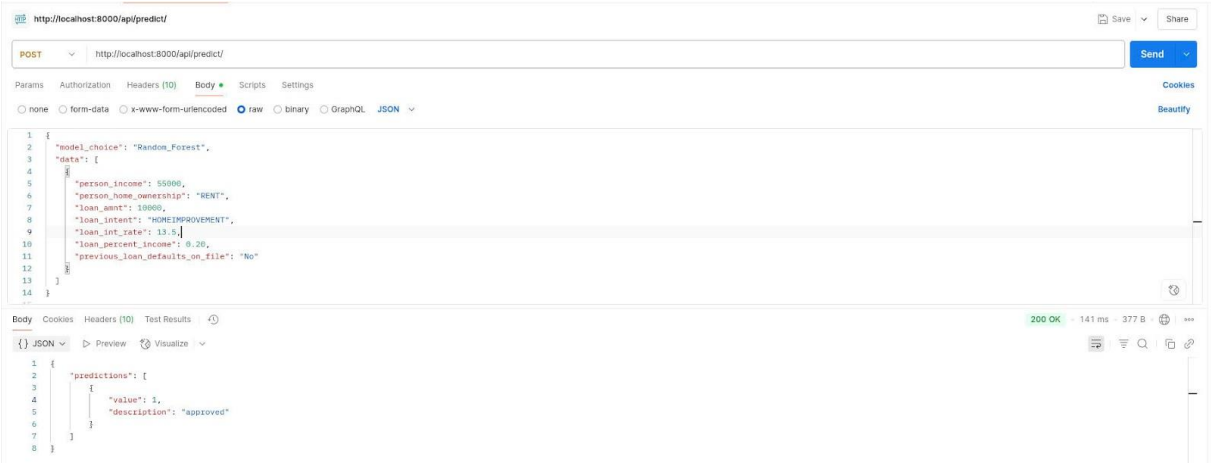
previous_loan_defaults_on_file	String	Yes, No	Previous default history

Output sẽ bao gồm giá trị (value) và mô tả kết quả dự đoán (description) ở dạng JSON:

{"predictions": [{ "value": <0/1>, "description": "approved/rejected" }] }

7.3 Demo

Thông qua API, chúng tôi sẽ demo quá trình gửi thông tin và nhận kết quả. API nhận đầu vào là thông tin tài chính của người vay và trả về kết quả dự đoán khoản vay có được chấp thuận hay không. Kết quả sẽ được hiển thị như hình ảnh:



Hình 7.2 Kết quả trả về thành công

Các trường hợp trả kết quả lỗi sẽ được trình bày trong phần phụ lục hình ảnh.

8. KẾT LUẬN

Qua quá trình thực hiện đồ án, chúng tôi đã vận dụng được các kiến thức chuyên ngành các môn. Chúng tôi đã thực hiện tiền xử lý dữ liệu, thăm dò sơ bộ để đánh giá dữ liệu và thực hiện trực quan để tìm kiếm thông tin giúp chúng tôi và mọi người có cái nhìn về nhiều góc độ hơn đối với vấn đề của bài toán. Sau khi áp dụng các kiến thức về mô hình học máy để thực nghiệm, kết quả đã tìm ra được mô hình phù hợp với bài toán dự đoán tình hình phê duyệt khoản vay, mô hình học máy Random Forest với Accuracy 0.03 và Marco F1 Score 0.90. Ngoài ra, chúng tôi đã sử dụng Django với mục đích thông qua API có thể gửi thông tin và nhận kết quả dự đoán.

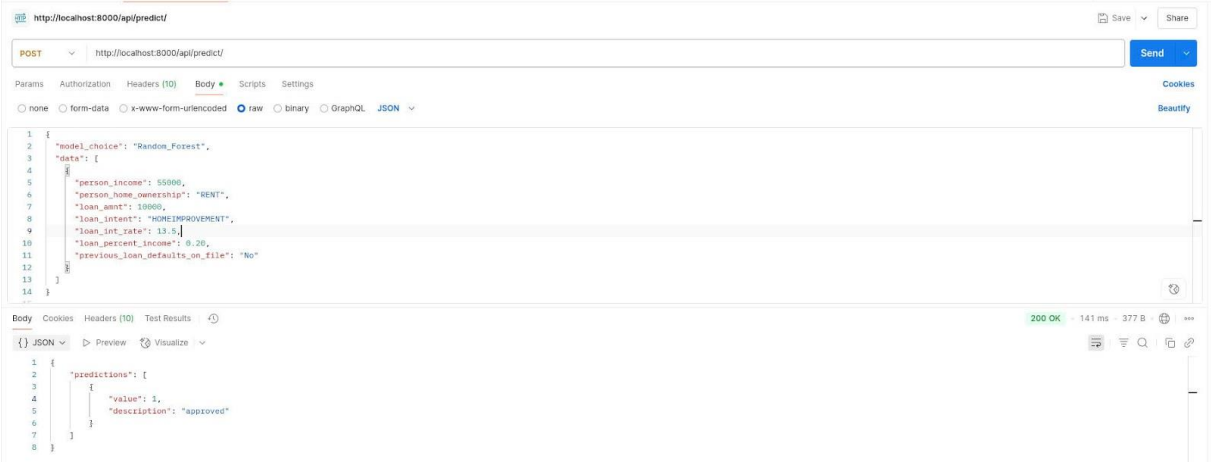
TÀI LIỆU THAM KHẢO

[1] Họ và tên tác giả: Tawei Lo. Link: https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data/data

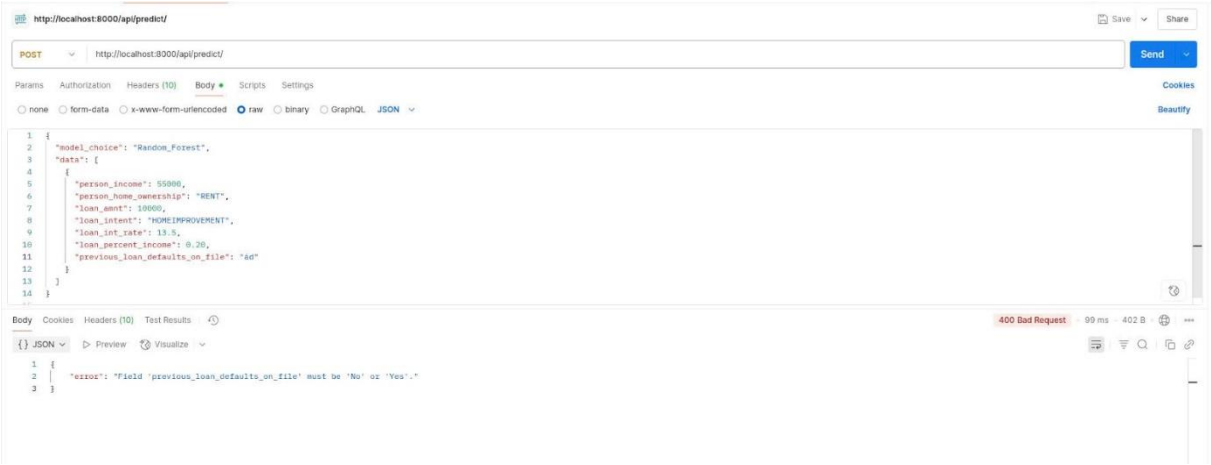
[2] Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition

[3] Báo cáo ngắn về input và output của API dự đoán. Link: curl -X POST -H "Content-Type: application/json" \ -H "X-API-Key:...

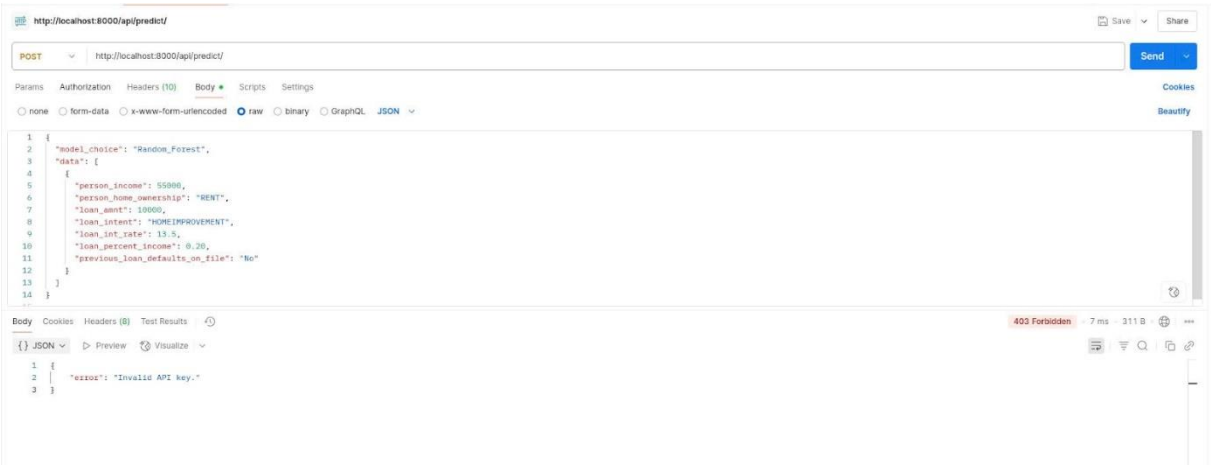
PHỤ LỤC HÌNH



Hình. Kết quả trả về thành công



Hình. Kết quả trả về error khi lỗi dữ liệu đầu vào



Hình. Kết quả trả về error khi API key không hợp lệ