**TRƯỜNG ĐẠI HỌC SƯ PHẠM HÀ NỘI**

**Khoa Công Nghệ Thông Tin**

**🙞🕮🙜**



**BÁO CÁO**

**CHUYÊN ĐỀ TỐT NGHIỆP KHOA HỌC** **DỮ LIỆU**

***Đề tài: Phân tích rủi ro tín dụng trong ngân hàng***

|  |  |  |
| --- | --- | --- |
| **Giảng viên** | **:** | Giang Thành Trung |
| **Nhóm sinh viên** | **:** | Đào Hải Nam - 715105158 |
|  |  | Nguyễn Đức Mậu - 715105146 |
| **Nhóm** | **:** | 11 |
| **Chủ đề** | **:** | 20 |

*Hà Nội, 2024*

MỤC LỤC

[Chương 1. Giới thiệu và chuẩn bị dữ liệu 2](#_Toc185261925)

[**1.1 Mục tiêu của dự án** 2](#_Toc185261926)

[**1.2 Phạm vi và giới hạn của dự án** 2](#_Toc185261927)

[**1.3 Tổng quan về dữ liệu** 4](#_Toc185261928)

[**1.4. Tiền xử lí dữ liệu** 5](#_Toc185261929)

[Chương 2. Khám phá dữ liệu và lựa chọn mô hình 10](#_Toc185261930)

[**2.1. Trực quan hóa dữ liệu** 10](#_Toc185261931)

[**2.2. Lựa chọn mô hình** 19](#_Toc185261932)

[Chương 3. Huấn luyện, tối ưu hóa và đánh giá mô hình 24](#_Toc185261933)

[**3.1 Mô hình sử dụng: RandomForestClassifier** 24](#_Toc185261934)

[Chương 4. Triển khai, ứng dụng và kết luận 30](#_Toc185261935)

[**4.1** **Triển khai và ứng dụng mô hình** 30](#_Toc185261936)

[**4.2 Theo dõi và cập nhật mô hình** 31](#_Toc185261937)

[***4.2.1 Theo dõi hiệu suất mô hình sau triển khai*** 31](#_Toc185261938)

[***4.2.2 Cập nhật mô hình khi cần thiết*** 31](#_Toc185261939)

[**4.3. Kết luận và đề xuất** 32](#_Toc185261940)

[***4.3.1 Tóm tắt kết quả chính*** 32](#_Toc185261941)

[***4.3.2 Đánh giá mục tiêu*** 32](#_Toc185261942)

[***4.3.3 Đề xuất cải tiến*** 32](#_Toc185261943)

[***4.3.4 Hạn chế và giải pháp*** 33](#_Toc185261944)

# Chương 1. Giới thiệu và chuẩn bị dữ liệu

## **1.1 Mục tiêu của dự án**

Mục tiêu chính của dự án: "Phân tích rủi ro tín dụng trong ngân hàng" là xây dựng và triển khai mô hình phân tích và dự đoán rủi ro tín dụng cho các khách hàng vay vốn. Điều này giúp ngân hàng xác định khả năng vỡ nợ của các khách hàng, từ đó tối ưu hóa quá trình ra quyết định cấp tín dụng, giảm thiểu rủi ro tài chính, và tối đa hóa lợi nhuận.

Trong lĩnh vực ngân hàng, việc xác định chính xác khả năng vỡ nợ của khách hàng vay là vô cùng cần thiết. Vấn đề này là một thách thức lớn do sự phức tạp của dữ liệu khách hàng, một dự đoán không chính xác có thể dẫn đến quyết định tín dụng sai lầm, gây tổn thất tài chính lớn cho ngân hàng.

## **1.2 Phạm vi và giới hạn của dự án**

***1.2.1. Phạm vi của dự án***

1.2.1.1. Phân tích dữ liệu

Dữ liệu: Sử dụng dữ liệu lịch sử tín dụng từ năm 2007 đến 2015, bao gồm thông tin tài chính và hành vi tín dụng của khách hàng.

Đặc trưng (Features): Dữ liệu bao gồm các đặc trưng như chính sách tín dụng, mục đích vay, lãi suất, tỷ lệ nợ trên thu nhập, điểm tín dụng FICO, thời gian sử dụng tín dụng, và các chỉ số tín dụng khác.

Mục tiêu: Phân tích dữ liệu để xác định các yếu tố ảnh hưởng đến rủi ro vỡ nợ và xây dựng mô hình dự đoán khả năng vỡ nợ của khách hàng.

* + - 1. Xây dựng mô hình

Phân tích và dự đoán khả năng tín dụng của người vay bằng cách sử dụng các kỹ thuật học máy như Hồi quy logistic hoặc Gradient Boosting. Các mô hình này rất cần thiết để giảm thiểu rủi ro tài chính trong ngân hàng.

* + 1. ***Giới hạn và giả định của dự án***
       1. Giới hạn

Giới hạn về thời gian: Dữ liệu chỉ từ năm 2007 đến 2015, không bao gồm các sự kiện kinh tế mới hơn hoặc các thay đổi chính sách sau năm 2015.

Giới hạn về dữ liệu: Dữ liệu có thể thiếu hoặc không phản ánh đầy đủ các yếu tố quan trọng khác như thay đổi chính sách kinh tế, các yếu tố cá nhân khác của khách hàng, hoặc tình trạng thị trường tài chính hiện tại.

Giới hạn về tính chính xác: Dữ liệu lịch sử có thể chứa lỗi hoặc không chính xác, đặc biệt là dữ liệu tự khai báo như thu nhập của khách hàng.

Giới hạn về mô hình: Mô hình học sâu có thể yêu cầu thời gian và tài nguyên tính toán lớn để huấn luyện và điều chỉnh, và kết quả có thể khó giải thích so với các mô hình đơn giản hơn.

* + 1. ***Giả định***

Giả định về tính đại diện: Dữ liệu từ năm 2007 đến 2015 được giả định là đủ đại diện cho xu hướng tín dụng và hành vi của khách hàng trong tương lai.

Giả định về mối quan hệ ổn định: Giả định rằng các mối quan hệ giữa các đặc trưng và khả năng vỡ nợ sẽ không thay đổi đáng kể theo thời gian, đảm bảo rằng mô hình có thể áp dụng cho dữ liệu mới.

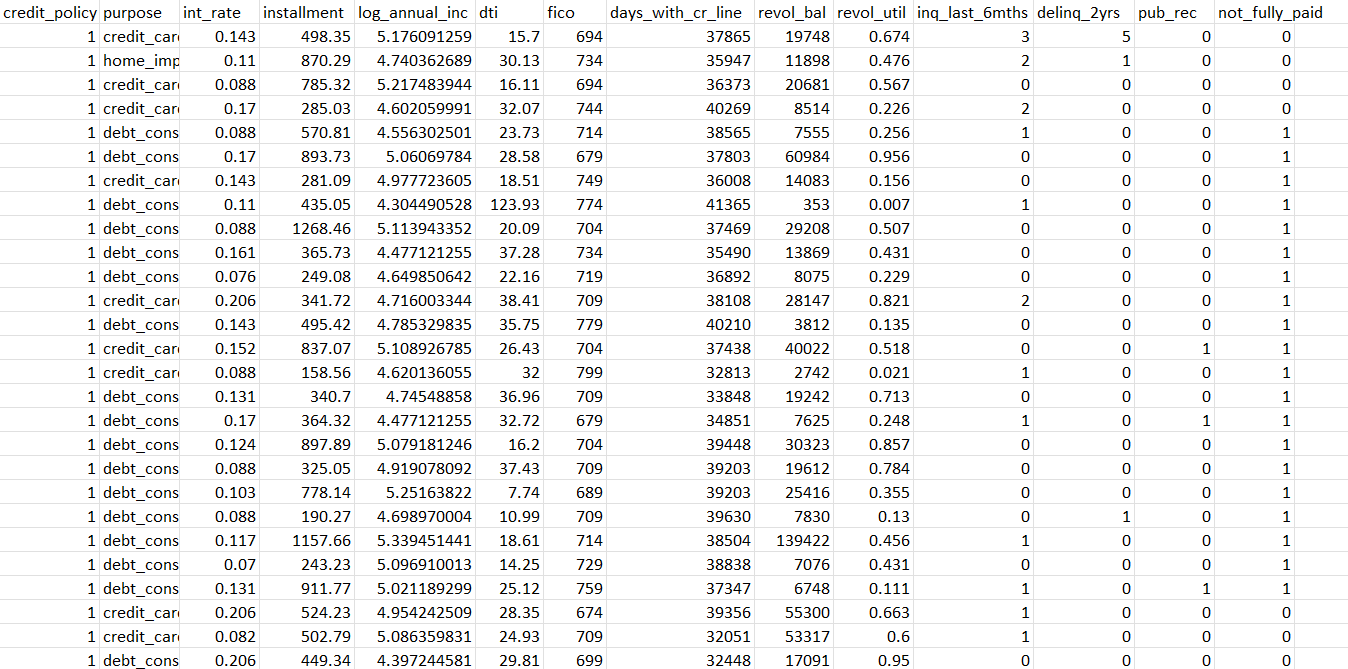
Giả định về tính độc lập: Giả định rằng các đặc trưng được sử dụng trong mô hình là độc lập hoặc đã được xử lý để giảm thiểu sự phụ thuộc lẫn nhau.

Giả định về môi trường kinh tế: Giả định rằng không có thay đổi lớn trong chính sách tín dụng hoặc các yếu tố kinh tế vĩ mô có thể làm thay đổi hành vi của khách hàng và ảnh hưởng đến kết quả mô hình.

Việc xác định rõ phạm vi và hiểu rõ các giới hạn, giả định sẽ giúp dự án thực hiện hiệu quả hơn và đưa ra những dự báo thực tế, đồng thời quản lý được các rủi ro tiềm tàng trong quá trình phân tích và dự đoán.

## **1.3 Tổng quan về dữ liệu**

**Dataset sử dụng: Lending club loan data**

****

**Cột dữ liệu trong datasets**

**credit.policy:** 1 nếu khách hàng đáp ứng tiêu chí bảo lãnh tín dụng của LendingClub.com, và 0 nếu không.

**purpose:** Mục đích của khoản vay (có các giá trị như "credit\_card", "debt\_consolidation", "educational", "major\_purchase", "small\_business", và "all\_other").

**int.rate:** Lãi suất của khoản vay, dưới dạng tỷ lệ phần trăm (một tỷ lệ 11% sẽ được lưu trữ dưới dạng 0.11). Những người vay được LendingClub.com đánh giá là rủi ro hơn sẽ bị áp dụng lãi suất cao hơn.

**installment:** Số tiền trả góp hàng tháng mà người vay phải trả nếu khoản vay được tài trợ.

**log.annual.inc:** Logarit tự nhiên của thu nhập hàng năm do người vay tự khai báo.

**dti:** Tỷ lệ nợ trên thu nhập của người vay (số tiền nợ chia cho thu nhập hàng năm).

**Fico:** Điểm tín dụng FICO của người vay.

**days.with.cr.line:** Số ngày mà người vay đã có hạn mức tín dụng.

**revol.bal:** Số dư quay vòng của người vay (số tiền chưa thanh toán vào cuối chu kỳ thanh toán thẻ tín dụng).

**revol.util:** Tỷ lệ sử dụng hạn mức tín dụng quay vòng của người vay (số tiền đã sử dụng so với tổng hạn mức tín dụng có sẵn).

**inq.last.6mths:** Số lượng yêu cầu của người vay trong 6 tháng qua.

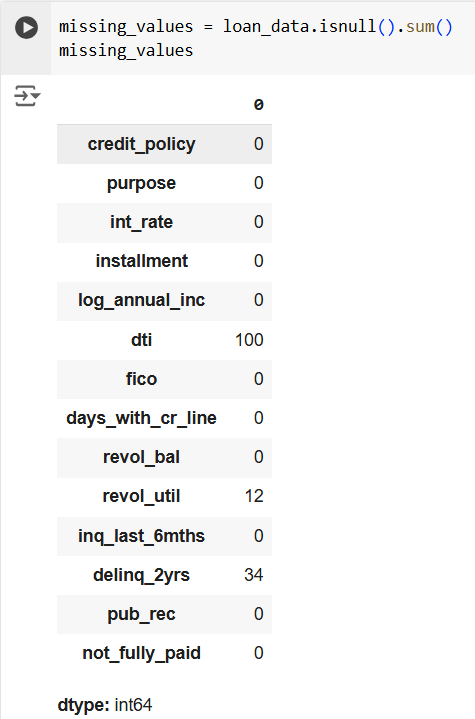
**delinq.2yrs:** Số lần người vay đã quá hạn thanh toán trên 30 ngày trong 2 năm qua.

**pub.rec:** Số lượng hồ sơ công khai tiêu cực của người vay (các vụ phá sản, tịch biên thuế, hoặc phán quyết).

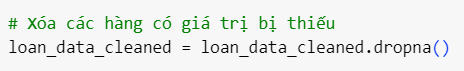
## **1.4. Tiền xử lí dữ liệu**

**Số bản ghi là 37066**

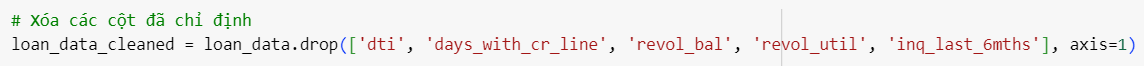
**Kiểm tra các giá trị thiếu**

****

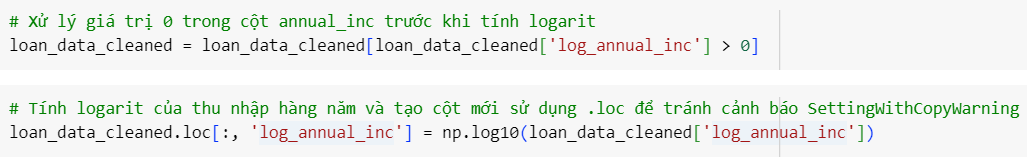
**Xử lí loại bỏ các giá trị thiếu**

****

**Xóa các cột chỉ định**

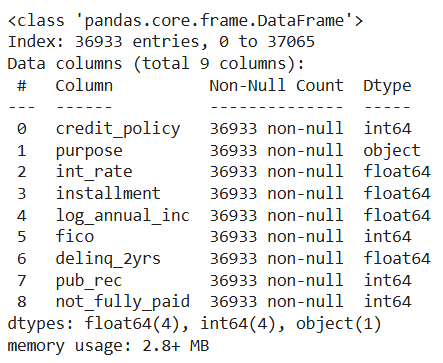
****

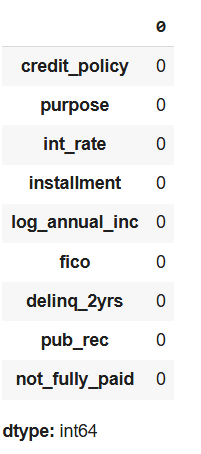
**Xử lí giá trị 0 và tính logarit thu nhập hằng năm**

****

**Kiểm tra lại thông tin, các giá trị thiếu sau khi làm sạch**

**Số bản ghi là 36933**

****

****

**Không còn dữ liệu null**

**Mã hóa dữ liệu phân loại**

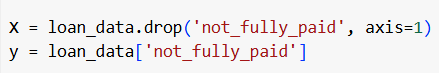


Sử dụng LabelEncoding() chuyển đổi dữ liệu dạng chuỗi sang dạng object, cụ thể là cột ‘purpose’.

**fit:** Tìm tất cả các giá trị duy nhất trong cột 'purpose' và ánh xạ mỗi giá trị chuỗi sang một số nguyên.

**transform:** Chuyển đổi các giá trị trong cột 'purpose' thành các số nguyên đã được ánh xạ ở bước trước.

**Tách dữ liệu đầu vào**

****

Tách dữ liệu thành biến đầu vào (features) và biến mục tiêu (target).

Biến X lấy toàn bộ dữ liệu trừ cột not\_fully\_paid để sử dụng làm biến đầu vào.

Biến y lấy riêng cột not\_fully\_paid để làm biến mục tiêu.

Tách riêng features và target giúp dễ dàng áp dụng các thuật toán học máy.

Biến X được chuẩn hóa hoặc xử lý thêm, trong khi biến y giữ nguyên.

**Xử lí giá trị thiếu**

****

Sử dụng SimpleImputer xử lí giá trị thiếu (NaN) trong tập dữ liệu X bằng cách thay thế các giá trị thiếu trong mỗi cột bằng giá trị trung bình của cột đó.

**fit(X):** Phân tích dữ liệu trong X và tính toán các tham số cần thiết (ở đây là trung bình của mỗi cột).

**transform(X):** Thay thế tất cả các giá trị thiếu (NaN) trong mỗi cột bằng trung bình đã tính ở bước fit.

**X\_imputed:** mảng đã được xử lý, tất cả kết quả NaN đã được thay thế.

# Chương 2. Khám phá dữ liệu và lựa chọn mô hình

## **2.1. Trực quan hóa dữ liệu**

***2.1.1. Báo cáo khám phá dữ liệu từ trực quan hóa***

2.1.1.1. Mục tiêu phân tích

Báo cáo nhằm trình bày kết quả khám phá được từ biểu đồ trực quan hóa dữ liệu khoản vay, bao gồm tỷ lệ vỡ nợ, phân tích các yếu tố tín dụng, và nhận diện các rủi ro tiềm ẩn liên quan đến các nhóm khách hàng khác nhau.

2.1.1.2. Dữ liệu và công cụ sử dụng

**Nguồn dữ liệu:** Dữ liệu lịch sử khoản vay từ 2007 đến 2015 của LendingClub.com

**Các biến số chính:**

**credit\_policy:** tiêu chuẩn tín dụng

**purpose:** mục đích vay

**int\_rate:** lãi suất khoản vay

**fico:** điểm tín dụng FICO

**not\_fully\_paid:** biến phân loại vỡ nợ (0: vỡ nợ, 1: không vỡ nợ)

**delinq.2yrs:** Số lần người vay đã quá hạn thanh toán trên 30 ngày trong 2 năm qua.

**log.annual.inc:** Logarit tự nhiên của thu nhập hàng năm do người vay tự khai báo.

**Công cụ sử dụng**: Power BI.

* + 1. ***Kết quả khám phá từ việc trực quan hóa***

2.1.2.1 Phân tích về khách hàng vay tín dụng

Tổng số lượng khách hàng vay tín dụng là 37066, với sự phân bố đa dạng về mục đích vay và mức độ tín dụng.

Lượng khách hàng đáp ứng tiêu chí tín dụng (**credit\_policy)** là 37066, chiếm 100% tổng số.

Số lượng hồ sơ tiêu cực là 4386, chiếm 12%.

Kết luận:

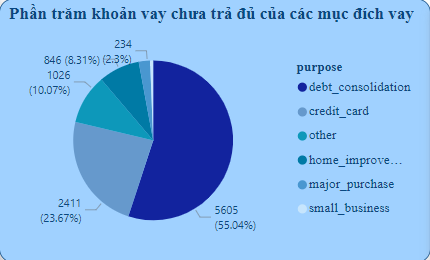
Các yếu tố tín dụng như credit.policy, purpose và pub.rec đóng vai trò quan trọng trong việc đánh giá rủi ro tín dụng.

Tăng cường tiêu chí đánh giá đối với khách hàng không đáp ứng đủ tiêu chí tín dụng.

Theo dõi sát sao khách hàng có hồ sơ tiêu cực để giảm thiểu rủi ro vỡ nợ.

Đưa ra biện pháp cải thiện chất lượng tín dụng cho nhóm khách hàng có rủi ro cao.

2.1.2.2. Phân tích tỷ lệ vỡ nợ theo mục đích vay



**Mục tiêu phân tích:**

Phân tích tỷ lệ vỡ nợ của các khoản vay theo mục đích vay, dựa trên dữ liệu từ biểu đồ Pie Chart trong Power BI.

**Dữ liệu và phương pháp:**

Sử dụng biểu đồ pie chart với mục đích (purpose) của khoản vay và tổng của not.fully.paid (0: vỡ nợ; 1: không vỡ nợ).

Ý nghĩa: tỷ lệ vỡ nợ của mục đích vay cho thấy nhóm khách hàng nào có rủi ro cao.

**Kết quả phân tích:**

Debt\_consolidation (hợp nhất nợ) có tỷ lệ vỡ nợ cao nhất (chiếm 55.04%).

Small\_business (mua sắm lớn) có tỷ lệ vỡ nợ thấp nhất (chiếm 0.61%).

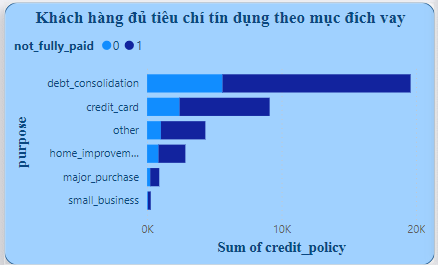
Các khoản vay với mục đích debt\_consolidation với credit\_card có tỷ lệ vỡ nợ cao hơn so với các mục đích khác, đây có thể là dấu hiệu cảnh báo rủi ro tín dụng.

**Kết luận và khuyến nghị:**

Các mục đích vay như debt\_consolidation, other hay credit\_card cần được kiểm soát chặt chẽ hơn với các điều kiện tín dụng nghiêm ngặt để giảm thiểu rủi ro vỡ nợ.

Cần có chính sách giám sát và hỗ trợ khách hàng vay với mục đích như debt\_consolidation hay credit\_card để giảm tỷ lệ vỡ nợ.

2.1.2.3. Phân tích số lượng khách hàng đáp ứng tiêu chí tín dụng theo mục đích vay

****

**Mục tiêu phân tích:**

Thống kê số lượng khách hàng đáp ứng tiêu chí tín dụng theo mục đích vay dựa trên sự phân loại về tỷ lệ vỡ nợ.

**Phương pháp phân tích:**

Sử dụng biểu đồ Stacked Bar Chart với cột Y là purpose (mục đích), cột X là sum of credit.policy và trường phân loại legend là not.fully.paid (0: vỡ nợ, 1: không vỡ nợ).

**Kết quả trực quan hóa:**

**Debt\_consolidation(hợp nhất nợ):**

Có lượng khách hàng cao, cho thấy nhiều khách hàng đáp ứng tiêu chí tín dụng lựa chọn vay với mục đích hợp nhất nợ.

Tuy nhiên số lượng khoản vay không đầy đủ chiếm tỷ lệ đáng kể (chiếm 41%) cho thấy sự rủi ro tiềm ẩn của nhóm này.

**Credit\_card (Thẻ tín dụng):**

Số lượng khoản vay cao và một phần không nhỏ trong số đó gặp phải tình trạng vỡ nợ.

Điều này thể hiện rằng nhu cầu sử dụng thẻ tín dụng có rủi ro cao và cần được theo dõi chặt chẽ hơn.

**Home\_improvement (Cải thiện nhà):**

Dù lượng khách hàng không cao thế nhưng tỷ lệ vỡ nợ lại lớn hơn so với những mục đích kể trên (43%).

Điều này cho thấy các khoản vay với mục đích cải thiện nhà có tỷ lệ vỡ nợ cao.

**Small\_business(kinh doanh nhỏ) và Major Purchase (Mua sắm lớn):**

Lượng khách hàng tương đối thấp, nhưng vẫn có sự xuất hiện của vỡ nợ.

Những khoản vay này có tính rủi ro trung bình, cần thêm nhiều yếu tố để đánh giá.

**Kết luận và khuyến nghị**

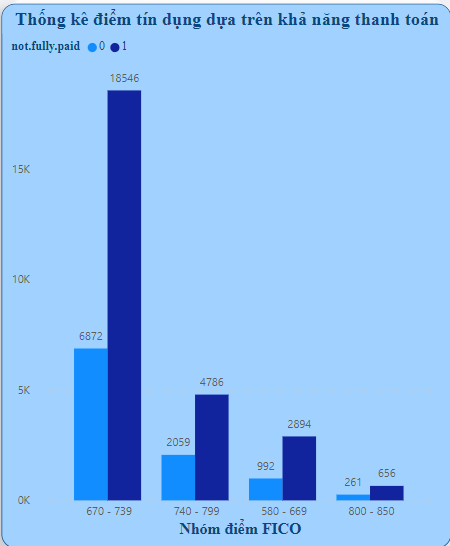
Ngay cả khi khách hàng đáp ứng đủ tiêu chí tín dụng, vẫn xảy ra rủi ro vỡ nợ. Các nhóm debt\_consolidation và credit\_card có xu hướng vỡ nợ cao, dù cho khách hàng đủ tiêu chí tín dụng.

Cần tập trung vào các khoản vay với mục đích debt\_consolidation và credit\_card để tìm hiểu nguyên nhân và đưa ra các biện pháp giải quyết

Có thể cân nhắc nâng cao tiêu chuẩn đối với nhóm khách hàng có rủi ro vỡ nợ cao để giảm tỷ lệ thất thoát

Tăng cường giám sát sau khi cấp vốn để phát hiện sớm dấu hiệu rủi ro trong quá trình thanh toán

2.1.2.4 Thống kê điểm tín dụng FICO dựa trên not.fully.paid (chưa trả hết nợ)

****

**Mục tiêu phân tích:**

Giúp phân tích mối quan hệ giữa điểm FICO của khách hàng và not.fully.paid (chưa trả hết nợ), giúp cho xác định nhóm điểm FICO có nguy cơ vỡ nợ cao, qua đó hỗ trợ xây dựng chiến lược tín dụng phù hợp và giảm thiểu rủi ro

**Phương pháp phân tích:**

Sử dụng biểu đồ Clustered Column Chart với trục X là điểm FICO của khách hàng được phân thành các nhóm (580 – 669, 670 – 739, …), trục Y là số lượng khách hàng tương ứng với nhóm điểm FICO và legend là not.fully.paid để phân loại theo trạng thái vỡ nợ (0: vỡ nợ; 1: không vỡ nợ)

**Kết quả trực quan hóa**

**Nhóm FICO trong khoảng từ 580 - 669:**

Nhóm **580 - 669** có tỷ lệ vỡ nợ cao hơn đáng kể so với các nhóm khác, với 992 trường hợp vỡ nợ và 2,894 trường hợp không vỡ nợ.

Rủi ro tín dụng cao, với tỷ lệ vỡ nợ tương đối lớn. Điều này có thể do khả năng tài chính yếu hoặc lịch sử tín dụng kém.

Siết chặt kiểm soát tín dụng: Yêu cầu các tài liệu chứng minh thu nhập hoặc tài sản thế chấp trước khi phê duyệt khoản vay.

Chính sách lãi suất cao hơn để bù đắp rủi ro.

Hỗ trợ cải thiện điểm FICO: Cung cấp các gói tư vấn tài chính giúp khách hàng quản lý nợ và tăng điểm tín dụng.

**Nhóm FICO trong khoảng từ 670 - 739:**

Nhóm **670 - 739** có số lượng khách hàng cao nhất, cả ở trạng thái vỡ nợ (6,872) và không vỡ nợ (18,546).

Mặc dù số lượng khách hàng lớn, tỷ lệ vỡ nợ vẫn hiện diện, cho thấy đây là nhóm có mức độ rủi ro trung bình.

Giám sát kỹ lưỡng: Phân tích kỹ lịch sử tín dụng và khả năng thanh toán trước khi cấp khoản vay lớn.

Khuyến khích thanh toán sớm: Cung cấp ưu đãi cho những khách hàng duy trì lịch sử trả nợ tốt.

**Nhóm FICO trong khoảng từ 800 - 850:**

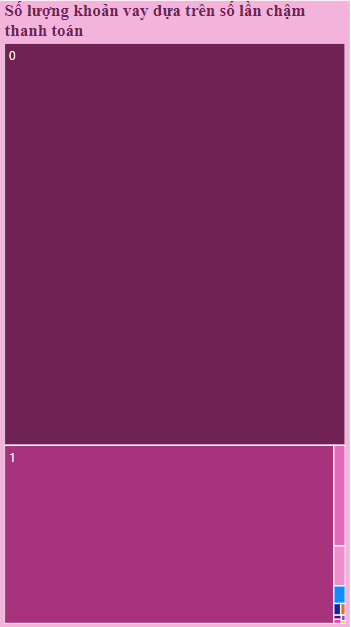
Nhóm **800 - 850** có ít khách hàng nhất và tỷ lệ vỡ nợ rất thấp, chỉ với 261 trường hợp vỡ nợ.

Rủi ro thấp nhất, phần lớn khách hàng trong nhóm này không vỡ nợ, chứng tỏ sự đáng tin cậy tài chính.

Tạo ưu đãi tài chính hấp dẫn: Các chương trình ưu đãi lãi suất thấp để giữ chân khách hàng lâu dài.

Tăng cường dịch vụ khách hàng để duy trì lòng tin và mối quan hệ lâu dài

2.1.2.5 Phân tích số lượng khoản vay theo số lần chậm thanh toán

****

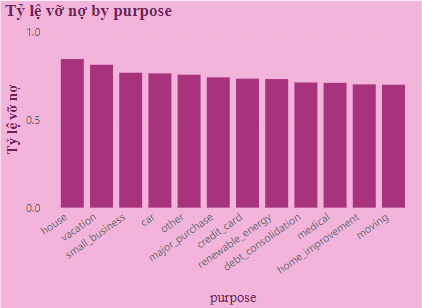
**Mục tiêu chính**: Phân tích mối quan hệ giữa số lần quá hạn trong 2 năm (delinq.2yrs) và chính sách tín dụng (credit.policy), nhằm xác định phân bố số lượng khoản vay dựa trên lịch sử quá hạn của khách hàng.

**Phương pháp**: Sử dụng biểu đồ treemap với delinq.2yrs làm danh mục (category) và Count of credit.policy làm giá trị (value), để hiển thị số lượng khoản vay theo từng mức độ quá hạn.

**Kết quả**: Các ô lớn hơn trong treemap cho thấy phần lớn khoản vay thuộc về nhóm khách hàng không có hoặc có rất ít lần quá hạn, trong khi các nhóm có nhiều lần quá hạn có số lượng khoản vay ít hơn.

**Kết luận**: Phần lớn khoản vay tập trung vào nhóm khách hàng có lịch sử tín dụng tốt (ít hoặc không quá hạn), cho thấy chính sách tín dụng ưu tiên nhóm này.

2.1.2.6 Phân tích tỷ lệ vỡ nợ theo mục đích vay

****

**Mục tiêu chính**: Phân tích **tỷ lệ vỡ nợ** (khoản vay không được trả đủ) cho từng mục đích vay (purpose), nhằm xác định những mục đích nào có nguy cơ vỡ nợ cao hơn.

**Phương pháp**:

Trục X: Các mục đích vay tiền (purpose) được chia thành các nhóm như house, vacation, small\_business, car, debt\_consolidation, medical, v.v.

Trục Y: Tỷ lệ vỡ nợ (từ 0 đến 1), biểu thị mức độ rủi ro của khách hàng theo từng mục đích vay.

Mỗi cột đại diện cho một mục đích vay và tỷ lệ vỡ nợ tương ứng.

**Kết quả:**

Nhóm house, vacation, small\_business có tỷ lệ vỡ nợ cao nhất (xấp xỉ 1.0), nghĩa là rủi ro rất lớn khi cấp khoản vay cho các mục đích này.

Nhóm medical, home\_improvement, moving có tỷ lệ vỡ nợ thấp hơn, chứng tỏ đây là những mục đích vay tương đối an toàn.

Các mục đích khác như debt\_consolidation, major\_purchase, car nằm ở mức trung bình.

**Kết luận:**

**Mục đích vay** với tỷ lệ vỡ nợ cao hơn (các cột cao) có thể cần chính sách tín dụng khắt khe hơn hoặc cần được theo dõi kỹ càng hơn.

**Mục đích vay** với tỷ lệ vỡ nợ thấp có thể được ưu tiên và có thể giảm rủi ro cho tổ chức tài chính.

## **2.2. Lựa chọn mô hình**

***2.2.1. Trình bày mô hình và thuật toán được chọn để phân tích dữ liệu***

Mô hình sử dụng: Random Forest Classification

**Random Forest Classification** là một thuật toán học máy dựa trên kỹ thuật ensemble learning, sử dụng nhiều cây quyết định (decision trees) để đưa ra kết quả phân loại cuối cùng thông qua cơ chế bỏ phiếu đa số. Các đặc điểm chính của Random Forest bao gồm:

Xử lý dữ liệu phi tuyến: Random Forest không yêu cầu dữ liệu phải tuyến tính, do đó phù hợp với các bài toán phức tạp như dự đoán rủi ro tín dụng, nơi các mối quan hệ giữa biến đầu vào và đầu ra thường không tuyến tính.

Khả năng kháng nhiễu: Thuật toán ít bị ảnh hưởng bởi dữ liệu bị mất cân bằng hoặc các giá trị ngoại lai.

Tính ổn định: Với cơ chế tổng hợp từ nhiều cây quyết định, mô hình này ít bị overfitting hơn so với một cây đơn lẻ.

Giải thích được: Có thể sử dụng tính năng "feature importance" để đánh giá tầm quan trọng của các biến đầu vào đối với kết quả dự đoán.

Mô hình sử dụng: Logistic Regression

**Logistic Regression** là một thuật toán học máy thuộc nhóm tuyến tính, được sử dụng phổ biến trong các bài toán phân loại nhị phân (binary classification). Mặc dù tên gọi "regression" (hồi quy), Logistic Regression thực chất là một mô hình phân loại, chứ không phải mô hình hồi quy như Linear Regression.

***2.2.2. Lý do chọn mô hình***

**Hiệu quả trong các bài toán phân loại rủi ro tín dụng:**

**Random Forest**

Random Forest đã được chứng minh là hiệu quả trong các bài toán về tài chính và tín dụng vì:

Khả năng phân biệt tốt giữa khách hàng rủi ro cao và rủi ro thấp nhờ việc kết hợp nhiều cây quyết định.

Độ chính xác cao và khả năng tổng quát hóa tốt khi dữ liệu có nhiều biến đầu vào với các mức độ quan trọng khác nhau.

**Đặc tính của dữ liệu:**

Dữ liệu tín dụng thường chứa cả biến định lượng (như thu nhập, số dư tài khoản) và định tính (như lịch sử tín dụng, nghề nghiệp). Random Forest có khả năng xử lý tốt cả hai loại dữ liệu này.

Mô hình này cũng không yêu cầu nhiều về tiền xử lý dữ liệu, ví dụ như chuẩn hóa hoặc chuyển đổi biến.

**Logistic Regression**

Đơn giản và hiệu quả:

Tính dễ hiểu: Logistic Regression là một trong những mô hình phân loại cơ bản và dễ giải thích. Mối quan hệ giữa các đặc trưng đầu vào và xác suất thuộc lớp có thể được diễn giải trực tiếp từ các trọng số của mô hình.

Hiệu quả với dữ liệu nhỏ: Khi số lượng dữ liệu không quá lớn và không quá phức tạp, Logistic Regression có thể đạt hiệu quả cao mà không cần nhiều tài nguyên tính toán.

Phù hợp với bài toán tín dụng:

Xử lý tốt các bài toán phân loại nhị phân: Dự đoán rủi ro tín dụng thường là bài toán nhị phân (khách hàng rủi ro hoặc không rủi ro), phù hợp với Logistic Regression.

Đánh giá xác suất: Logistic Regression không chỉ phân loại mà còn cung cấp xác suất cho từng dự đoán, giúp hỗ trợ đưa ra quyết định rủi ro (ví dụ: thiết lập ngưỡng phê duyệt tín dụng).

***2.2.3 Quá trình chia dữ liệu thành tập huấn luyện, tập kiểm tra***

**Chia dữ liệu thành tập huấn luyện (training sets) và tập kiểm tra (datasets)**

Sử dụng hàm train\_test\_split từ thư viện scikit-learn để chia dữ liệu đầu vào thành hai tập riêng biệt:

Tập huấn luyện (X\_train, y\_train): dữ liệu được dùng để huấn luyện mô hình.

Tập kiểm tra (X\_test, y\_test): Dữ liệu được dữ lại để đánh giá hiệu suất của mô hình trên dữ liệu chưa từng thấy trước đó.

Các tham số đầu vào:

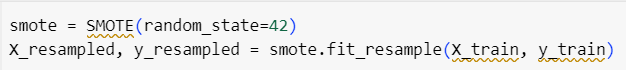
**X\_imputed:** Tập dữ liệu đặc trưng (features) đã được xử lý để loại bỏ giá trị thiếu (NaN). Kích thước ban đầu: (36919, 13).

**y:** Biến mục tiêu (cột not\_fully\_paid), kích thước (36919,).

**test\_size=0.2:** 20% dữ liệu được đưa vào tập X\_test và y\_test. Phần còn lại (80%) đưa vào X\_train và y\_train.

**random\_state=42:** giá trị này giúp chia dữ liệu cùng một cách mỗi lần chạy, đảm bảo kết quả ổn đỉnh khi huấn luyện mô hình. Nếu không đặt, dữ liệu sẽ được ngẫu nhiên khác nhau.

**Sử dụng SMOTE để giải quyết mất cân bằng dữ liệu**



**smote = SMOTE(random\_state=42)**

Tạo một đối tượng SMOTE với tham số ‘random\_state=42’ thiết lập cho việc tái lập kết quả. Điều này đảm bảo kết quả nhất quán mỗi khi chạy mã.

**X\_resampled, y\_resampled = smote.fit\_resample(X\_train, y\_train)**

Áp dụng SMOTE để tạo thêm dữ liệu từ lớp thiếu số.

Đầu vào:

**X\_train:** tập dữ liệu đặc trưng (features) trước khi cân bằng.

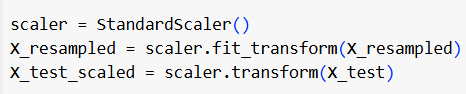
**y\_train:** tập nhãn (labels) tương ứng trước khi cân bằng.

Đầu ra:

**X\_resampled:** dữ liệu đặc trưng sau khi thêm các mẫu tổng hợp.

**Y\_resampled:** dữ liệu nhãn tương ứng sau khi cân bằng.

**Chuẩn hóa dữ liệu sử dụng StandardScaler()**



Thực hiện chuẩn hóa dữ liệu nhằm làm cho các đặc trưng trong dữ liệu có trung bình bằng không và độ lệch chuẩn bằng 1 giúp cho mô hình học máy hoạt động hiểu quả hơn.

**X\_resampled = scaler.fit\_transform(X\_resampled)**

fit: Tính trung bình (μ) và độ lệch chuẩn (σ) cho từng cột dữ liệu của X\_resampled.

transform: áp dụng công thức để chuẩn hóa từng giá trị trong X\_resampled.

**X\_test\_scaled = scaler.transform(X\_test)**

Sử dụng trung bình và độ lệch chuẩn đã được tính từ tập huấn luyện X\_resampled.

Áp dụng công thức chuẩn hóa tương tự cho X\_test.

# Chương 3. Huấn luyện, tối ưu hóa và đánh giá mô hình

## **3.1 Mô hình sử dụng: RandomForestClassifier**

***3.1.1 Huấn luyện mô hình***

Quá trình huấn luyện mô hình

Khởi tạo mô hình Random Forest

****

Random Forest Classifier là mô hình học máy được xây dựng bằng cách kết hợp nhiều cây quyết định (**Decision Trees**) để đưa ra dự đoán.

Mỗi cây được huấn luyện trên các mẫu khác nhau của dữ liệu và chỉ sử dụng một tập con các đặc trưng, giúp tăng cường tính đa dạng và giảm overfitting.

Tham số random\_state=42: Đảm bảo kết quả có thể tái hiện bằng cách cố định trạng thái ngẫu nhiên.

Dữ liệu đầu vào

****

X\_resampled: Dữ liệu được huấn luyện đã được chuẩn hóa sau khi áp dụng SMOTE

y\_resampled: Nhãn của tập huấn luyện, đã được cân bằng bởi SMOTE.

fit: Huấn luyện mô hình Random Forest bằng cách học các mẫu từ dữ liệu.

Dự đoán trên tập kiểm tra

****

**predict()**

Dùng để thực hiện dự đoán trên tập dữ liệu đầu vào (X\_test\_scaled) dựa trên mô hình Random Forest đã được huấn luyện (rf\_model).

Kết quả dự đoán là nhãn phân loại (label).

Kết quả sẽ là:

0: Nếu dự đoán mẫu thuộc lớp rủi ro tín dụng.

1: Nếu dự đoán mẫu không thuộc lớp rủi ro tín dụng.

Đầu vào:

X\_test\_scaled:

Dữ liệu đã được kiểm tra trước đo bằng StandardScaler.

Mỗi hàng đại diện cho một mẫu, mỗi cột là một đặc trưng.

Đầu ra:

y\_pred:

Mảng chứa các nhãn dự đoán cho tập kiểm tra (X\_test\_scaled).

****

**predict\_proba():**

Tính toán xác suất dự đoán của mỗi mẫu thuộc từng lớp trong bài toán phân loại.

Kết quả trả về một mảng 2D, trong đó mỗi hàng là một mẫu, mỗi cột tương ứng với xác suất thuộc từng lớp

Cột 0: xác suất thuộc lớp 0 (rủi ro tín dụng), cột 1: xác suất thuộc lớp 1 (không rủi ro tín dụng)

[:, 1]: Lấy xác suất thuộc lớp 1 (cột thứ 1) từ kết quả của predict\_proba().

Đầu vào:

X\_test\_scaled: dữ liệu kiểm tra đã được chuẩn hóa, giống như dòng đầu tiên

Đầu ra:

y\_prob:

Mảng 1D chứa xác suất dự đoán cho từng mẫu thuộc lớp rủi ro tín dụng

* + 1. ***Đánh giá hiệu suất mô hình***

****

**accuracy\_score():**

Tính độ chính xác tổng quát (Accuracy) của mô hình

Kết quả: Accuracy: 0.8465601300108342

Mô hình dự đoán đúng ~84.66% số mẫu trong tập kiểm tra

****

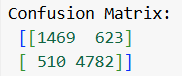
**Confusion\_matrix():**

Tính ma trận nhầm lẫn (Confusion Matrix) để đánh giá kết quả phân loại

Ma trận này cho biết số lượng dự đoán đúng và sai của mô hình theo từng lớp

Đầu vào là y\_test: nhãn thực tế từ tập kiểm tra, y\_pred: Nhãn dự đoán từ mô hình

Kết quả:

****

True Positive (TP) = 1469:

Mô hình dự đoán chính xác rủi ro tín dụng cho 1469 khách hàng. Đây là kết quả mong muốn, giúp ngân hàng tránh cho vay đến những khách hàng có nguy cơ vỡ nợ.

False Negative (FN) = 623:

Mô hình dự đoán không rủi ro tín dụng, nhưng thực tế khách hàng này có rủi ro tín dụng. Những trường hợp này có thể gây tổn thất lớn vì mô hình không nhận diện được các khách hàng rủi ro.

False Positive (FP) = 510:

Mô hình dự đoán rủi ro tín dụng, nhưng thực tế các khách hàng này không rủi ro tín dụng. Đây là dự đoán sai, có thể làm ngân hàng từ chối những khách hàng đáng tin cậy, dẫn đến mất cơ hội kinh doanh.

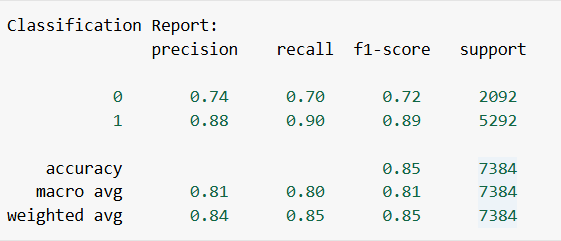
True Negative (TN) = 4782:

Mô hình dự đoán chính xác không rủi ro tín dụng cho 4782 khách hàng. Đây là kết quả tốt, giúp ngân hàng đưa ra quyết định đúng đắn.

****

**classification\_report():** Tạo một bảng báo cáo tóm tắt các chỉ số đánh giá hiệu năng phân loại của mô hình

Kết quả:



#### **Lớp 0 (Rủi ro tín dụng):**

#### **Precision = 0.74 (74%)**:

Trong số các khách hàng mà mô hình dự đoán là **rủi ro tín dụng, 74%** là chính xác.

Mô hình vẫn nhầm lẫn một phần (26%) các khách hàng không rủi ro tín dụng thành rủi ro tín dụng (**FP**).

**Recall = 0.70 (70%)**:

Trong số các khách hàng thực sự **rủi ro tín dụng**, mô hình chỉ nhận diện đúng được **70%**.

**30%** khách hàng rủi ro tín dụng bị bỏ sót **(FN).**

**F1-Score = 0.72 (72%)**:

Kết quả này cho thấy hiệu suất tổng thể cho lớp rủi ro tín dụng là trung bình, cần cải thiện thêm.

#### **Lớp 1 (Không rủi ro tín dụng):**

**Precision = 0.88 (88%)**:

Trong số các khách hàng mà mô hình dự đoán là **không rủi ro tín dụng**, **88%** là chính xác.

**12%** các khách hàng rủi ro tín dụng bị nhầm là không rủi ro tín dụng (**FN**).

**Recall = 0.90 (90%)**:

Trong số các khách hàng thực sự **không rủi ro tín dụng,** mô hình nhận diện đúng **90%**.

**10%** khách hàng không rủi ro tín dụng bị nhầm thành rủi ro tín dụng (**FP**).

**F1-Score = 0.89 (89%)**:

Hiệu suất tổng thể cho lớp không rủi ro tín dụng rất tốt.

Hiệu suất tổng quan:

**Accuracy = 85%:**

Tỷ lệ dự đoán đúng trên toàn bộ tập kiểm tra. Mô hình đang hoạt động khá tốt với độ chính xác cao.

**Macro Average:**

**Precision = 0.81, Recall = 0.80, F1-Score = 0.81:**

Trung bình các chỉ số này được tính đều giữa hai lớp. Lớp 1 (không rủi ro tín dụng) có hiệu suất tốt hơn kéo trung bình này lên.

**Weighted Average:**

**Precision = 0.84, Recall = 0.85, F1-Score = 0.85:**

Trung bình các chỉ số này được tính có trọng số dựa trên số lượng mẫu trong mỗi lớp. Vì số lượng khách hàng không rủi ro tín dụng (lớp 1) chiếm ưu thế, nên hiệu suất của lớp này ảnh hưởng mạnh đến chỉ số tổng thể.

# Chương 4. Triển khai, ứng dụng và kết luận

## **Triển khai và ứng dụng mô hình**

* + 1. ***Quy trình triển khai mô hình***

**Chuẩn bị môi trường triển khai:** Cài đặt các thư viện cần thiết như scikit-learn, joblib, pandas, numpy, và công cụ streamlit để xây dựng API hoặc giao diện người dùng.

Xây dựng hệ thống giao tiếp: Xây dựng giao diện người dùng bằng Streamlit, cho phép nhập liệu trực tiếp và hiển thị kết quả dự đoán.

Triển khai mô hình: Tải mô hình đã được huấn luyện từ file đã lưu, Xử lý dữ liệu đầu vào, chuẩn hóa và thực hiện dự đoán.

Các bước kiểm tra mô hình sau khi triển khai

Kiểm thử các chức năng.

Đảm bảo các dữ liệu đầu vào sai định dạng được xử lý phù hợp.

Kiểm tra giao diện với bộ dữ liệu thử nghiệm.

Kiểm thử hiệu suất: Đánh giá thời gian xử lý cho từng yêu cầu dự đoán.

So sánh dự đoán với thực tế: Đối chiếu kết quả dự đoán với dữ liệu thực tế sau một khoảng thời gian để đánh giá độ chính xác.

* + 1. ***Ứng dụng mô hình vào giải quyết vấn đề thực tế***

Ví dụ cụ thể:

Ngân hàng: Dự đoán khả năng không trả nợ của khách hàng và tối ưu hóa quyết định cấp tín dụng.

Công ty tài chính: Gửi thông báo cảnh báo đến bộ phận quản lý rủi ro nếu khách hàng nằm trong nhóm nguy cơ cao.

## **4.2 Theo dõi và cập nhật mô hình**

### ***4.2.1 Theo dõi hiệu suất mô hình sau triển khai***

**Thu thập dữ liệu thực tế**:

Liên tục thu thập các trường hợp dự đoán sai hoặc không chính xác (e.g., dự đoán khách hàng sẽ trả nhưng không trả).

So sánh phân phối dữ liệu mới với dữ liệu huấn luyện ban đầu.

**Đánh giá hiệu suất định kỳ**:

Sử dụng các chỉ số như Accuracy, Precision, Recall, F1 Score, và AUC-ROC trên dữ liệu thực tế.

**Giám sát rủi ro**:

Theo dõi tỷ lệ quyết định dựa trên mô hình dẫn đến kết quả không mong muốn.

Lập báo cáo định kỳ để đánh giá tính hiệu quả của mô hình trong thực tế.

### ***4.2.2 Cập nhật mô hình khi cần thiết***

**Xử lý dữ liệu mới**:

Gộp dữ liệu mới với dữ liệu cũ và xây dựng một tập huấn luyện cập nhật.

Đảm bảo các giá trị mới trong cột purpose hoặc các thay đổi khác được xử lý đúng cách.

**Re-train mô hình**:

Huấn luyện lại mô hình khi nhận thấy hiệu suất giảm đáng kể.

Sử dụng các kỹ thuật như transfer learning nếu chỉ muốn cập nhật một phần của mô hình.

**Kiểm thử lại**:

Lặp lại quy trình kiểm tra sau khi cập nhật mô hình.

Đảm bảo rằng mô hình mới không làm giảm hiệu suất trên các trường hợp đã hoạt động tốt trước đó.

## **4.3. Kết luận và đề xuất**

### ***4.3.1 Tóm tắt kết quả chính***

Mô hình đã đạt được độ chính xác khá cao trên tập kiểm tra (Accuracy = 85%).

Triển khai thành công trong môi trường giả lập và đã tích hợp với hệ thống quyết định tín dụng.

### ***4.3.2 Đánh giá mục tiêu***

**Mục tiêu ban đầu**: Xây dựng mô hình dự đoán khả năng không trả nợ.

**Mức độ đạt được**: Mô hình đã dự đoán chính xác phần lớn các trường hợp nguy cơ cao, hỗ trợ giảm rủi ro tín dụng.

### ***4.3.3 Đề xuất cải tiến***

**Tăng cường thu thập dữ liệu mới**:

Mở rộng tập dữ liệu để bao phủ các trường hợp

**Sử dụng thêm các kỹ thuật khác**:

Áp dụng mô hình Deep Learning nếu dữ liệu trở nên lớn hơn.

**Cải thiện giao diện người dùng**:

Tích hợp hệ thống với các dashboard trực quan (e.g., Power BI hoặc Tableau) để theo dõi kết quả dự đoán.

### ***4.3.4 Hạn chế và giải pháp***

**Hạn chế**:

Phân phối dữ liệu mất cân bằng (imbalance) có thể dẫn đến đánh giá sai mức độ rủi ro.

Mô hình hiện tại có thể không thích nghi tốt với dữ liệu mới nếu không cập nhật kịp thời.

**Giải pháp**:

Sử dụng kỹ thuật oversampling hoặc undersampling phù hợp hơn (SMOTE).

Theo dõi dữ liệu thực tế thường xuyên và lập lịch cập nhật mô hình định kỳ.