**TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN**

**KHOA TOÁN KINH TẾ**

# BÀI TIỂU LUẬN QUẢN TRỊ RỦI RO 2

## **ĐỀ TÀI:** MỘT SỐ PHƯƠNG PHÁP THỐNG KÊ CHO CHẤM ĐIỂM TÍN DỤNG VÀ DỰ BÁO KHẢ NĂNG VỠ NỢ

Họ và tên: Vũ Lê Anh Thư

Mã sinh viên: 11226148

Lớp: Toán kinh tế 64

Giảng viên: Ts. Nguyễn Thị Liên

Hà Nội, 2025

# MỤC LỤC


**1. GIỚI THIỆU** ..... 2

**1.1 Lý do nghiên cứu** ..... 2

**1.2 Mục tiêu nghiên cứu** ..... 2

**1.3 Tầm quan trọng của nghiên cứu** ..... 2

**1.4 Phạm vi nghiên cứu** ..... 2

**2. CỞ SỞ LÝ THUYẾT** ..... 3

**2.1 Rủi ro tín dụng** ..... 3

**2.2 Các phương pháp dự báo và chấm điểm tín dụng** ..... 3

***2.2.1 Hồi quy Logistic (Logistic Regression)*** ..... 3

***2.2.2 Mô hình Random Forest*** ..... 4

***2.2.3 Mô hình K-Nearest Neighbors (KNN)*** ..... 6

***2.2.4 Chấm điểm tín dụng (Credit Score)*** ..... 7

***2.2.5 Information Value (IV) và Weight of Evidence (WOE)*** ..... 8

**3. DỮ LIỆU** ..... 10

**3.1 Nguồn:** ..... 10

**3.2 Phân tích dữ liệu** ..... 10

**3.3 Xử lý và làm sạch dữ liệu** ..... 10

**3.4 Kiểm tra tình trạng cân bằng của dữ liệu** ..... 11

**4 KẾT QUẢ** ..... 12

**4.1 Hồi quy Logistic với biến gốc** ..... 12

**4.2 Hồi quy Logistic với biến WOE và chọn biến dựa trên InformationValue** ..... 12

**4.3 Mô hình Random Forest** ..... 17

**4.4 Mô hình K-Nearest Neighbors (KNN)** ..... 18

**5 KẾT LUẬN** ..... 19

**TÀI LIỆU THAM KHẢO** ..... 20

# 1. GIỚI THIỆU

## 1.1 Lý do nghiên cứu

Trong bối cảnh ngành tài chính – ngân hàng không ngừng phát triển, việc kiểm soát rủi ro tín dụng trở thành một yếu tố then chốt để đảm bảo hoạt động ổn định và hiệu quả của các tổ chức tín dụng. Rủi ro tín dụng xảy ra khi khách hàng không thực hiện nghĩa vụ trả nợ đúng hạn, gây thiệt hại đáng kể về tài chính. Dù đã có nhiều phương pháp được áp dụng trong việc đánh giá rủi ro, song vẫn còn tồn tại những hạn chế như: chưa tối ưu hóa việc khai thác dữ liệu, chưa ứng dụng đầy đủ các kỹ thuật học máy hiện đại hay chưa nâng cao được khả năng dự báo. Điều này đặt ra nhu cầu cấp thiết cho việc nghiên cứu và áp dụng các mô hình phân tích dữ liệu tiên tiến nhằm cải thiện hiệu quả dự đoán rủi ro và chấm điểm tín dụng.

## 1.2 Mục tiêu nghiên cứu

Nghiên cứu này hướng đến việc hỗ trợ các tổ chức tín dụng đưa ra quyết định cho vay chính xác hơn bằng cách ứng dụng các kỹ thuật phân tích dữ liệu và mô hình dự báo. Mục tiêu cụ thể bao gồm:

- Phân tích đặc điểm khách hàng để xác định những yếu tố có ảnh hưởng lớn đến khả năng trả nợ đúng hạn.
- Ứng dụng các mô hình dự đoán nhằm ước lượng xác suất khách hàng không thanh toán đúng kỳ hạn.
- So sánh hiệu quả của các phương pháp dự báo khác nhau để tìm ra mô hình tối ưu cho việc đánh giá rủi ro tín dụng.
- Đề xuất cách xây dựng hệ thống điểm tín dụng (credit scoring) từ mô hình, giúp ngân hàng dễ dàng phân loại khách hàng và đưa ra quyết định cho vay phù hợp.

## 1.3 Tầm quan trọng của nghiên cứu

Rủi ro tín dụng là một trong những nguyên nhân chính gây thiệt hại tài chính cho các ngân hàng và tổ chức tín dụng. Nếu không đánh giá chính xác khả năng trả nợ của khách hàng, tổ chức tín dụng có thể đưa ra các quyết định cho vay sai lệch, dẫn đến nợ xấu, ảnh hưởng đến lợi nhuận và uy tín.

Trong bối cảnh dữ liệu khách hàng ngày càng phong phú, việc ứng dụng các mô hình phân tích dữ liệu và học máy sẽ giúp ngân hàng khai thác thông tin hiệu quả hơn, dự đoán rủi ro sát thực tế hơn thay vì chỉ dựa vào kinh nghiệm hoặc đánh giá chủ quan. Từ đó, nâng cao khả năng quản lý rủi ro tín dụng và tối ưu hóa quyết định cho vay.

## 1.4 Phạm vi nghiên cứu

Bài luận sử dụng bộ dữ liệu khách hàng về rủi ro tín dụng từ Kaggle bằng công cụ R để xây dựng và đánh giá các mô hình dự báo rủi ro tín dụng và tính điểm tín dụng. Các mô hình nghiên cứu bao gồm hồi quy logistic với biến gốc, hồi quy logistic với biến đã chuyển đổi WOE, Random

Forest và K-Nearest Neighbors (KNN). Mô hình được đánh giá qua các chỉ số thống kê như AUC, độ chính xác và độ nhạy. Bài luận tập trung vào dự báo khả năng vỡ nợ của khách hàng, không mở rộng đến các yếu tố bên ngoài như biến động kinh tế hay chính sách tín dụng.

## 2. CỞ SỞ LÝ THUYẾT

### 2.1 Rủi ro tín dụng

Rủi ro tín dụng là rủi ro mà các tổ chức tài chính (như ngân hàng, công ty tài chính) phải đối mặt khi cho vay tiền hoặc cấp tín dụng, và có khả năng người vay hoặc đối tác tài chính không thể trả lại số tiền đã vay đúng hạn hoặc hoàn trả toàn bộ. Nói cách khác, đây là khả năng mà người vay hoặc đối tác tài chính không thực hiện được nghĩa vụ thanh toán nợ, dẫn đến tổn thất tài chính cho tổ chức cho vay.

Các loại rủi ro tín dụng gồm:

- Rủi ro vỡ nợ: Người vay không trả được nợ.
- Rủi ro tái cấu trúc nợ: Điều kiện nợ bị thay đổi.
- Rủi ro không thanh toán: Người vay không thanh toán dù có khả năng.

Quản lý rủi ro tín dụng bao gồm việc đánh giá tín dụng, yêu cầu tài sản đảm bảo và phân bổ rủi ro giữa các khoản vay khác nhau.

### 2.2 Các phương pháp dự báo và chấm điểm tín dụng

#### *2.2.1 Hồi quy Logistic (Logistic Regression)*

Hồi quy Logistic là một mô hình phân loại nhị phân, giúp dự báo xác suất mà một sự kiện xảy ra. Trong bài toán rủi ro tín dụng, mô hình này được sử dụng để dự đoán khả năng vỡ nợ (biến mục tiêu là "default. payment.next.month").

Công thức hồi quy Logistic:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}}$$

Trong đó:

- $P(Y = 1|X)$ là xác suất xảy ra sự kiện (khách hang vỡ nợ)
- $\beta_0, \beta_1, \dots, \beta_n$là các hệ số của mô hình, được ước lượng từ dữ liệu
- $X_1, X_2, \dots, X_n$ là các biến độc lập (các đặc điểm của khách hàng ảnh hướng tới khả năng trả nợ như số dư tín dụng, tình trạng thanh toán, tuổi, thu nhập, v.v)

Ta dự báo:

- Y = 1 (Có vỡ nợ) nếu $P(Y = 1) \geq 0.5$, tương đương:

$$\frac{1}{1 + e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \geq 0.5$$
$$\Leftrightarrow \beta_0 + \beta_1X_1 + \cdots + \beta_nX_n \geq 0$$

- Y = 0 (Không vỡ nợ) nếu $P(Y = 1) < 0.5$, tương đương:

$$\frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}} < 0.5$$
$$\Leftrightarrow \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n < 0$$

**Ưu điểm của mô hình Logistic:**

Mô hình logistic regression rất dễ hiểu vì các hệ số trong mô hình có thể được giải thích theo dạng log odds (logarithm của tỷ lệ xác suất). Ví dụ, nếu hệ số của một biến độc lập là 0.5, thì mỗi đơn vị tăng của biến này sẽ làm tỷ lệ log-odds của sự kiện tăng lên 0.5. Điều này giúp mô hình rất dễ áp dụng và kiểm tra bởi những người không chuyên về thống kê hay học máy.

**Nhược điểm của mô hình Logistic:**

Mô hình logit rất nhạy cảm với outliers (dữ liệu ngoại lai). Nếu có những điểm dữ liệu cực đoan (ví dụ: các giá trị rất lớn hoặc rất nhỏ so với phần còn lại), chúng có thể làm lệch kết quả của mô hình. Vì logistic regression tính toán xác suất dựa trên sự kết hợp của các yếu tố, những điểm dữ liệu bất thường này có thể ảnh hưởng mạnh mẽ đến các hệ số ước lượng của mô hình. Khi đó, kết quả của mô hình có thể không phản ánh chính xác thực tế, dẫn đến dự đoán sai cho các dữ liệu mới.

### *2.2.2 Mô hình Random Forest*

Random Forest là một thuật toán học máy mạnh mẽ thuộc nhóm học có giám sát (supervised learning), được phát triển bởi Leo Breiman vào năm 2001. Thuật toán này kết hợp nguyên lý "bagging" (bootstrap aggregating) và mô hình cây quyết định (Decision Tree) để tạo thành một "rừng" gồm nhiều cây khác nhau. Mục tiêu của Random Forest là cải thiện độ chính xác dự báo và giảm thiểu hiện tượng quá khớp (overfitting) mà các mô hình cây đơn lẻ thường gặp phải.

Điểm nổi bật của Random Forest là tính ngẫu nhiên trong cả quá trình lấy mẫu dữ liệu và chọn biến khi tách nhánh, giúp các cây thành phần đa dạng và ít tương quan với nhau. Sự kết hợp của nhiều cây "yếu" này, thông qua nguyên tắc bỏ phiếu (với phân loại) hoặc lấy trung bình (với hồi quy), tạo ra một mô hình tổng thể có độ chính xác cao và ổn định.

Nhờ những ưu điểm này, Random Forest được ứng dụng rộng rãi trong nhiều lĩnh vực như tài chính, y tế, phân tích rủi ro tín dụng, và khai thác dữ liệu lớn.

**Các thành phần của cây quyết định**

- Nút gốc (Root Node):

Đây là nút nằm ở đầu cây quyết định. Từ nút này, quá trình phân chia dữ liệu bắt đầu dựa trên các đặc trưng (tính năng) khác nhau.

- Nút quyết định (Decision Node):

Là các nút xuất hiện sau khi phân chia từ nút gốc. Tại đây, dữ liệu tiếp tục được chia nhỏ hơn nữa dựa trên các giá trị của biến đặc trưng.

- Nút lá (Leaf Node/Terminal Node):

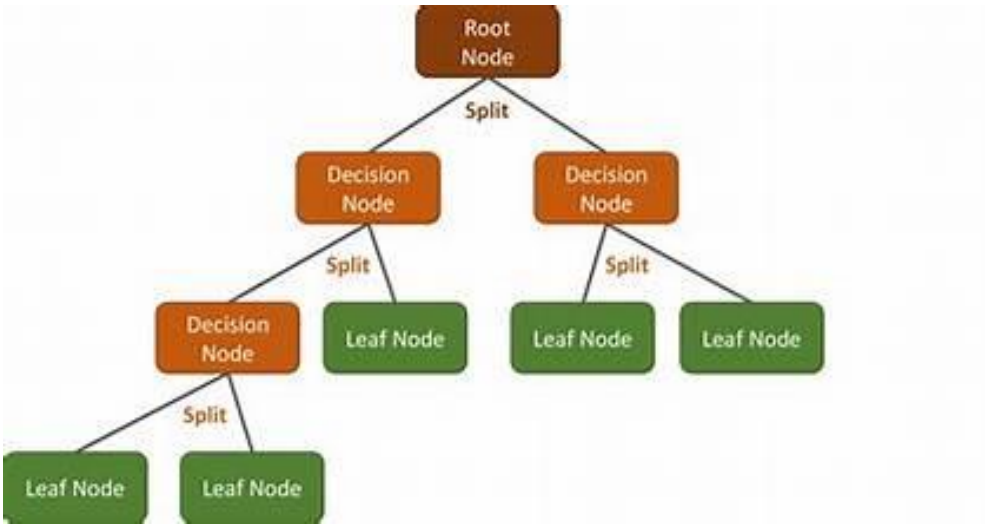
Là các nút cuối cùng trên cây, không còn phân chia thêm được nữa. Mỗi nút lá đại diện cho một kết quả hoặc một nhóm phân loại cụ thể.

- Cây con (Sub-tree):

Là một phần nhỏ của cây quyết định, giống như một nhánh con trong biểu đồ tổng thể. Mỗi cây con cũng có thể được coi là một cây quyết định nhỏ hơn.

- Tia bớt (Pruning):

Là quá trình cắt bớt một số nhánh hoặc nút của cây quyết định nhằm giảm độ phức tạp của cây, tránh tình trạng overfitting (quá khớp với dữ liệu huấn luyện).




*Hình 1: Minh họa thuật toán Random Forest*

**Nguyên lý hoạt động:**

*Bước 1: Chọn số lượng cây*

Chọn T là số lượng cây thành phần sẽ được xây dựng trong mô hình.

*Bước 2: Chọn số lượng thuộc tính tại mỗi node*

Chọn m là số lượng các thuộc tính (biến) sẽ được dùng để phân chia tại mỗi node của cây.

Thông thường, m nhỏ hơn tổng số thuộc tính p rất nhiều và được giữ cố định trong suốt quá trình xây dựng cây.

*Bước 3: Dựng các cây quyết định*

Với mỗi cây:

a) Tạo tập mẫu bootstrap: Lấy mẫu ngẫu nhiên có hoàn lại n mẫu từ tập dữ liệu gốc.
b) Chia node: Ở mỗi node, chọn ngẫu nhiên m thuộc tính và chọn thuộc tính tốt nhất trong số đó để phân chia.\
c) Phát triển cây: Mỗi cây được phát triển lớn nhất có thể, không cắt tỉa (pruning).

*Bước 4: Tổng hợp kết quả*

Sau khi xây dựng xong T cây, để phân loại một đối tượng, thuật toán thu thập kết quả phân lớp của tất cả các cây và sử dụng kết quả được nhiều cây chọn nhất (đa số phiếu) làm kết quả cuối cùng.

Tỉ lệ lỗi tổng thể phụ thuộc vào độ mạnh của từng cây và mối quan hệ giữa các cây (tính đa dạng).

**Lưu ý về bootstrap và out-of-bag**

Khi lấy mẫu bootstrap cho mỗi cây, trung bình chỉ khoảng 2/3 số phần tử trong tập huấn luyện được chọn vào mẫu, còn khoảng 1/3 phần tử không nằm trong mẫu đó (out-of-bag).

Các phần tử out-of-bag này có thể dùng để đánh giá hiệu suất mô hình mà không cần tập kiểm tra riêng biệt.

**Ưu điểm của Random Forest**

- Độ chính xác cao: Random Forest thường cho kết quả dự đoán chính xác và ổn định nhờ kết hợp nhiều cây quyết định.
- Giảm overfitting: Nhờ kỹ thuật lấy mẫu bootstrap và chọn ngẫu nhiên biến tại mỗi node, mô hình giảm được hiện tượng overfitting so với cây quyết định đơn lẻ.
- Khả năng xử lý dữ liệu lớn: Hiệu quả với các bộ dữ liệu có nhiều biến, nhiều quan sát.

**Nhược điểm của Random Forest:**

- Kết quả chậm hơn: Do thuật toán xây dựng nhiều cây, điều này làm tăng độ phức tạp và thời gian tính toán, đặc biệt khi xây dựng hàng trăm hoặc hàng nghìn cây. Vì vậy, nó có thể không hiệu quả đối với các dự đoán thời gian thực.
- Yêu cầu nhiều tài nguyên hơn: Vì Random Forest xử lý tập dữ liệu lớn hơn, nó cần nhiều tài nguyên hơn để lưu trữ dữ liệu.
- Phức tạp hơn: Dự đoán của một cây quyết định đơn lẻ dễ diễn giải hơn so với một rừng nhiều cây.

### *2.2.3 Mô hình K-Nearest Neighbors (KNN)*

K-Nearest Neighbors (KNN) là một thuật toán học máy giám sát (supervised learning) dùng để phân loại hoặc dự đoán giá trị liên tục dựa trên sự gần gũi của các điểm dữ liệu. KNN là thuật

toán phi tham số (non-parametric), tức là không giả định một mô hình cụ thể cho dữ liệu, mà sử dụng đặc tính gần nhất của các điểm trong không gian tính toán để đưa ra dự đoán.

**Cách thức hoạt động của KNN:**

*Bước 1: Xác định số k*

Chọn một số k (số lượng láng giềng gần nhất) trước khi bắt đầu phân loại.

*Bước 2: Tính khoảng cách*

Với mỗi điểm dữ liệu cần phân loại, tính khoảng cách giữa điểm đó và các điểm trong tập huấn luyện (thường sử dụng khoảng cách Euclidean).

Khoảng cách Euclidean giữa hai điểm $(x_1, x_2, \dots, x_n)$ và $(y_1, y_2, \dots, y_n)$ là:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

*Bước 3: Chọn k láng giềng gần nhất*

Lấy k điểm dữ liệu có khoảng cách nhỏ nhất với điểm cần phân loại.

*Bước 4: Dự đoán*

Phân loại (Classification): Dựa vào sự đa số của các láng giềng (k láng giềng có nhãn giống nhau). Nếu đa số các láng giềng thuộc vào một lớp, thì điểm cần phân loại sẽ thuộc vào lớp đó.

Hồi quy (Regression): Dự đoán giá trị liên tục cho điểm dữ liệu mới bằng cách tính giá trị trung bình (hoặc có thể là trung bình có trọng số) của các giá trị mục tiêu của k láng giềng gần nhất.

**Ưu điểm của KNN:**

- Đơn giản và dễ hiểu: KNN rất dễ triển khai và dễ hiểu, không yêu cầu mô hình phức tạp.
- Không cần giả định về phân phối dữ liệu: KNN không yêu cầu giả định về phân phối dữ liệu, rất linh hoạt với các loại dữ liệu khác nhau.

**Nhược điểm của KNN:**

- Nhạy cảm với dữ liệu nhiễu: Các điểm dữ liệu nhiễu (outliers) có thể ảnh hưởng lớn đến kết quả dự đoán của KNN vì thuật toán phụ thuộc vào các láng giềng gần nhất.
- Khó khăn với không gian dữ liệu cao: KNN gặp khó khăn khi làm việc với dữ liệu có không gian chiều cao (high-dimensional data), vì khi số chiều tăng lên, khoảng cách giữa các điểm trở nên khó phân biệt rõ ràng (hiện tượng "curse of dimensionality").

### *2.2.4 Chấm điểm tín dụng (Credit Score)*

Chấm điểm tín dụng là một phương pháp tính toán nhằm đánh giá mức độ tín nhiệm của khách hàng. Mục đích là đưa ra một điểm số thể hiện khả năng thanh toán nợ của khách hàng, từ đó hỗ trợ các tổ chức tín dụng trong việc quyết định có cho vay hay không.

Trong các mô hình như hồi quy logistic, điểm tín dụng có thể được tính bằng cách sử dụng hàm xác suất từ mô hình logistic. Điểm tín dụng sẽ được tính bằng cách chuyển đổi xác suất dự đoán từ mô hình thành điểm số, thường có giá trị trong khoảng từ 300 đến 850.

Công thức tính điểm tín dụng:

$$Credit\ Score = Base\ Score + 100P(Y = 1|X)$$

Trong đó:

- $Credit\ Score$ là điểm tín dụng
- $P(Y = 1|X)$ là xác suất khách hàng vỡ nợ dự đoán từ mô hình hồi quy logistic
- $Base\ Score$ là điểm số cơ bản, được thiết lập tùy theo hệ thống chấm điểm

### *2.2.5 Information Value (IV) và Weight of Evidence (WOE)*

Information Value (IV) là một chỉ số đo lường mối quan hệ giữa biến độc lập và biến mục tiêu, giúp lựa chọn các biến quan trọng trong mô hình dự báo. IV được tính dựa trên sự khác biệt giữa tỷ lệ xác suất vỡ nợ và không vỡ nợ trong các nhóm giá trị của một biến.

Công thức tính IV:

$$IV = \sum_{i=1}^{k} \big(P(good_i) - P(bad_i)\big).\ln\frac{P(good_i)}{P(bad_i)}$$

Trong đó:

- $P(good_i)$ và $P(bad_i)$ là tỷ lệ xác suất của nhóm I trong các nhóm "good"(không vỡ nợ) và "bad" (vỡ nợ)
- $k$ là số lượng nhóm phân loại

| IV | Ý nghĩa |
|---|---|
| < 0.02 | Không hữu ích cho việc dự đoán |
| 0.02 - 0.1 | Yếu tố dự đoán yếu |
| 0.1 - 0.3 | Yếu tố dự đoán trung bình |
| 0.3 - 0.5 | Yếu tố dự đoán mạnh |
| > 0.5 | Đáng ngờ hoặc dự đoán quá tốt |

Weight of Evidence (WOE) là giá trị giúp chuyển đổi các biến định tính thành các giá trị số mà mô hình hồi quy logistic có thể sử dụng. WOE được tính bằng tỷ lệ odds (xác suất xảy ra và không xảy ra) của từng nhóm.

Công thức tính WOE:

$$WOE = \ln \frac{P(good_i)}{P(bad_i)}$$

Trong đó:

- $P(good_i)$ và $P(bad_i)$ là tỷ lệ xác suất của nhóm I trong các nhóm "good"(không vỡ nợ) và "bad" (vỡ nợ)

### 2.3 Các chỉ số đánh giá hiệu quả của mô hình

Độ chính xác (Accuracy): Là tỷ lệ dự đoán đúng của mô hình

Công thức:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Độ nhạy (Sensitivity): Tỷ lệ dự đoán đúng các trường hợp vỡ nợ.

Công thức:

$$Sensitivity = \frac{TP}{TP + FN}$$

Độ đặc hiệu (Specificity): Tỷ lệ dự đoán đúng các trường hợp không vỡ nợ.

Công thức:

$$Specificity = \frac{TN}{TN + FP}$$

Trong đó:

- TP: True Positive (dự đoán đúng là vỡ nợ).
- TN: True Negative (dự đoán đúng là không vỡ nợ).
- FP: False Positive (dự đoán sai là vỡ nợ).
- FN: False Negative (dự đoán sai là không vỡ nợ).

ROC (Receiver Operating Characteristic): Đường cong thể hiện sự trade-off giữa độ nhạy và độ đặc hiệu của mô hình.

AUC (Area Under Curve): Diện tích dưới đường cong ROC, cho biết khả năng phân biệt của mô hình giữa các lớp (vỡ nợ và không vỡ nợ).

AUC nằm trong khoảng từ 0 đến 1, càng gần 1 thì mô hình càng tốt.

- AUC = 1: Mô hình phân loại hoàn hảo.
- AUC = 0.5: Mô hình không phân biệt được, giống như việc đoán ngẫu nhiên.

- AUC < 0.5: Mô hình không tốt, và có thể cần phải cải thiện hoặc thử một mô hình khác.

## 3. DỮ LIỆU

### 3.1 Nguồn:

Bộ dữ liệu này được thu thập từ các khách hàng tín dụng của các ngân hàng tại Đài Loan trong khoảng thời gian từ tháng 4 năm 2005 đến tháng 9 năm 2005. Nó bao gồm thông tin về các yếu tố nhân khẩu học, lịch sử thanh toán, dữ liệu tín dụng, và các thông tin về hóa đơn thanh toán của khách hàng tín dụng. Dữ liệu này được công bố trên UCI Machine Learning Repository và có mục đích phân tích khả năng khách hàng sẽ bị vỡ nợ trong tháng tiếp theo dựa trên các thông tin có sẵn.

### 3.2 Phân tích dữ liệu

**Các biến trong bộ dữ liệu:**

- ID: Mã định danh khách hàng (tương ứng với mỗi khách hàng trong bộ dữ liệu).
- LIMIT_BAL: Số dư tín dụng được cấp (tính bằng đồng NT).
- SEX: Giới tính của khách hàng (1 = nam, 2 = nữ).
- EDUCATION: Trình độ học vấn của khách hàng (1 = học sau đại học, 2 = đại học, 3 = trung học phổ thông, 4 = khác, 5 = không rõ, 6 = không rõ).
- MARRIAGE: Tình trạng hôn nhân của khách hàng (1 = đã kết hôn, 2 = độc thân, 3 = khác).
- AGE: Tuổi của khách hàng.
- PAY_0 - PAY_6: Tình trạng trả nợ của khách hàng từng tháng từ tháng 9/2005 đổ về tháng 4/2005 (tính theo mức độ trễ hạn, từ -1 (trả đúng hạn) đến 9 (trễ hạn lớn hơn hoặc bằng 9 tháng).
- BILL_AMT1 - BILL_AMT6: Số tiền hóa đơn trong từng tháng từ tháng 9/2005 đổ về tháng 4/2005 (đơn vị: NT dollar).
- PAY_AMT1 - PAY_AMT6: Số tiền đã thanh toán trong từng tháng từ tháng 9/2005 đổ về tháng 4/2005 (đơn vị: NT dollar).
- default.payment.next.month: Trạng thái vỡ nợ trong tháng tiếp theo (1 = vỡ nợ, 0 = không vỡ nợ).

### 3.3 Xử lý và làm sạch dữ liệu

Trong quá trình xử lý dữ liệu, một số giá trị bất thường và không hợp lý đã được điều chỉnh để đảm bảo tính chính xác và thống nhất của bộ dữ liệu. Cụ thể, các bước xử lý dữ liệu được thực hiện như sau:

- Biến PAY_0 đến PAY_06 phản ánh tình trạng thanh toán trong 6 tháng của khách hàng. Các giá trị -2 và -1 (ngụ ý thanh toán đúng hạn) nên quy đổi hết về 0. Việc này giúp làm rõ tình trạng "đúng hạn" cho các khách hàng có trạng thái trả nợ tốt.

- Biến BILL_AMT1 đến BILL_AMT6 và PAY_AMT1 đến PAY_AMT6 thể hiện số dư hóa đơn và các khoản đã thanh toán tồn tại một số giá trị âm, điều này không phù hợp trong bối cảnh tài chính nên ta quy đổi các giá trị này về 0 nhằm đảm bảo tính hợp lý và tránh việc sai lệch trong các phân tích tiếp theo.
- Biến EDUCATION thể hiện trình độ học vấn xuất hiện một số giá trị không hợp lệ là 0 và giá trị không xác định là 5,6. Do đó ta chuyển đổi nó thành 4, đại diện cho trường hợp khác để đơn giản hóa và duy trì tính nhất quán trong bộ dữ liệu.
- Biến MARRIAGE thể hiện tình trạng hôn nhân có một số giá trị không hợp lệ là 0 nên ta quy đổi nó thành 3, đại diện cho tình trạng hôn nhân "khác".

Tiếp đến, ta kiểm tra dữ liệu bị thiếu trên toàn bộ bộ dữ liệu và không phát hiện thấy bất kỳ giá trị NA nào trong bộ dữ liệu, điều này có nghĩa là tất cả các biến đều có đủ giá trị dữ liệu nên ta không cần xử lý NA.

**Xử lý giá trị ngoại lai (Outliers)**

- Sử dụng phương pháp Interquartile Range (IQR) để phát hiện các giá trị ngoại lai. IQR được tính bằng hiệu số giữa phân vị thứ 3 (Q3) và phân vị thứ 1 (Q1), và các giá trị nằm ngoài phạm vi từ $Q1 - 1.5IQR$ đến $Q3 + 1.5IQR$ được xem là các giá trị ngoại lai.
- Sau khi phát hiện ra các giá trị ngoại lai, ta xử lý chúng bằng cách áp dụng phương pháp Winsorization cho các biến LIMIT_BAL và AGE. Phương pháp này là một kỹ thuật thay thế các giá trị ngoại lai bằng giá trị tại phân vị 1% hoặc 99% của dữ liệu.
- Đối với các biến BILL_AMT1 - BILL_AMT6 và PAY_AMT1 - PAY_AMT6, ta không xử lý các giá trị ngoại lai đối với các biến này vì bản chất của các biến tài chính này có thể chứa các giá trị lớn hợp lý trong một số tình huống. Ví dụ như có thể có những khách hàng có số tiền hóa đơn hoặc số tiền thanh toán lớn bất thường nhưng hợp lý, chẳng hạn như khách hàng có dư nợ cao hoặc thanh toán lớn vào một tháng nào đó. Việc xử lý các giá trị ngoại lai trong những trường hợp này có thể làm mất đi các thông tin quan trọng hoặc làm giảm tính chính xác của mô hình dự báo.

Tiếp đến, ta chuyển các biến SEX, EDUCATION, MARRIAGE và default.payment.next.month về dạng factor còn các biến PAY_0 đến PAY_6 ta dữ nguyên dạng numeric với 0 là trả đúng hạn, 1 là quá hạn 1 tháng, v.v.

### 3.4 Kiểm tra tình trạng cân bằng của dữ liệu

Để xác định tính cân bằng của dữ liệu, chúng ta cần kiểm tra tỷ lệ phân bổ giữa các lớp của biến mục tiêu default.payment.next.month

Sau khi kiểm tra phân bố của biến này, ta thu được kết quả sau:

- Số lượng khách hàng không vỡ nợ (0): 3892 (tương đương 77.85% tổng số khách hàng).
- Số lượng khách hàng vỡ nợ (1): 1107 (tương đương 22.14% tổng số khách hàng).

Tỷ lệ này có thể chấp nhận được trong các bài toán phân loại, với tỷ lệ 80/20 là một mức chấp nhận trong nhiều tình huống ở thực tế với các bài toán chấm điểm tín dụng.

# 4 KẾT QUẢ

## 4.1 Hồi quy Logistic với biến gốc

Ta chia dữ liệu theo tập train và set tỷ lệ 70:30 và chạy mô hình biến phụ thuộc default.payment.next.month theo tất cả các biến còn lại. Sau đó ta dung stepwise để chọn ra mô hình tốt nhất dựa trên tiêu chí AIC và p_value ≤ 0.5, cuối cùng mô hình được chọn bao gồm các biến sau: MARRIAGE, PAY_0, PAY_3, PAY_6, BILL_AMT3, BILL_AMT5, PAY_AMT1, và PAY_AMT2. Kết quả của mô hình hồi quy logistic có dạng như sau:

| Biến | Ước lượng (Estimate) | Sai số chuẩn (Std. Error) | Giá trị z (z value) | P-value |
|---|---|---|---|---|
| (Intercept) | -1.473 | 0.0801 | -18.385 | < 2e-16 *** |
| MARRIAGE2 | -0.2441 | 0.0892 | -2.737 | 0.006194 ** |
| MARRIAGE3 | -0.1864 | 0.3943 | -0.473 | 0.636379 |
| PAY_0 | 0.8428 | 0.0628 | 13.425 | < 2e-16 *** |
| PAY_3 | 0.1716 | 0.0634 | 2.707 | 0.006781 ** |
| PAY_6 | 0.2159 | 0.0669 | 3.227 | 0.001250 ** |
| BILL_AMT3 | 6.378e-06 | 1.804e-06 | 3.536 | 0.000406 ** |
| BILL_AMT5 | -7.258e-06 | 2.034e-06 | -3.568 | 0.000360 ** |
| PAY_AMT1 | -2.129e-05 | 7.446e-06 | -2.859 | 0.004248 ** |
| PAY_AMT2 | -2.037e-05 | 6.935e-06 | -2.938 | 0.003303 ** |

Ma trận nhầm lẫn:

| Dự đoán \ Thực tế | 0 | 1 |
|---|---|---|
| 0 | 1129 | 236 |
| 1 | 38 | 96 |

Đánh giá mô hình:

- Accuracy: Độ chính xác của mô hình là 81.72%, tức là mô hình dự đoán đúng 81.72% số trường hợp trên tập kiểm tra
- AUC: Diện tích dưới đường cong ROC (AUC) là 0.7272, cho thấy mô hình có khả năng phân loại tốt.
- Sensitivity: 71.57%, tức là phát hiện được 71.57% trường hợp mắc nợ, nhưng vẫn còn 28.43% bị bỏ sót.
- Specificity: 82.72%, tức là nhận diện chính xác 82.72% các trường hợp không mắc nợ.

## 4.2 Hồi quy Logistic với biến WOE và chọn biến dựa trên InformationValue

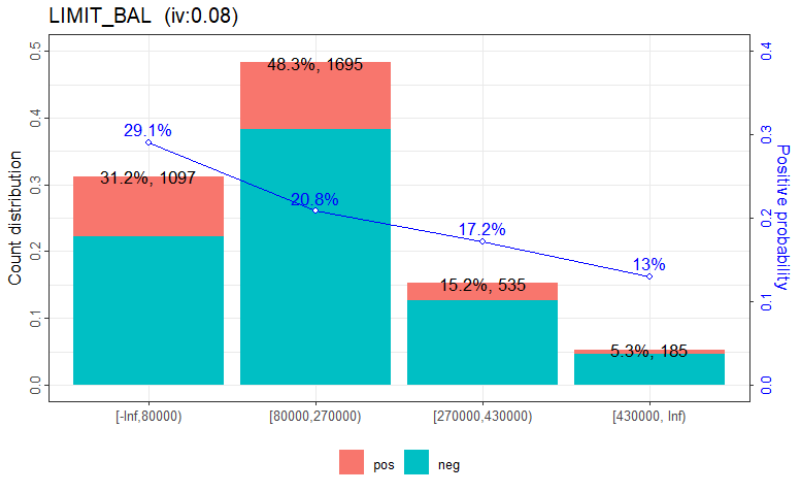
Để cải thiện mô hình hồi quy logistic, chúng tôi sử dụng phương pháp WOE để chuyển đổi các biến định tính và định lượng. Các bước xử lý dữ liệu bao gồm:

- Tính IV (Information Value) cho các biến độc lập để xác định mức độ ảnh hưởng của chúng đến biến mục tiêu được kết quả như sau:

| Variables | IV | Variables | IV | Variables | IV |
|---|---|---|---|---|---|
| **LIMIT_BAL** | 0.0725717514 | **PAY_6** | 0.1957439869 | **PAY_AMT4** | 0.0384351548 |
| **SEX** | 0.0009587721 | **BILL_AMT1** | 0.0120385887 | **PAY_AMT5** | 0.0251065989 |
| **EDUCATION** | 0.0254814240 | **BILL_AMT2** | 0.0074704816 | **PAY_AMT6** | 0.0465088019 |
| **MARRIAGE** | 0.0124405253 | **BILL_AMT3** | 0.0137875867 | | |
| **AGE** | 0.0248044846 | **BILL_AMT4** | 0.0136523622 | | |
| **PAY_0** | 0.7404526416 | **BILL_AMT5** | 0.0080768044 | | |
| **PAY_2** | 0.4677827813 | **BILL_AMT6** | 0.0184656435 | | |
| **PAY_3** | 0.3563237845 | **PAY_AMT1** | 0.0538699303 | | |
| **PAY_4** | 0.2703581616 | **PAY_AMT2** | 0.0591679935 | | |
| **PAY_5** | 0.2786892277 | **PAY_AMT3** | 0.0048977524 | | |

- Loại bỏ các biến có IV thấp (<0.02) nhằm giảm thiểu sự nhiễu loạn trong mô hình và ta được dữ liệu mới gồm các biến: LIMIT_BAL, EDUCATION, AGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5,  PAY_6, PAY_AMT1, PAY_AMT2, PAY_AMT4, PAY_AMT5, PAY_AMT6.

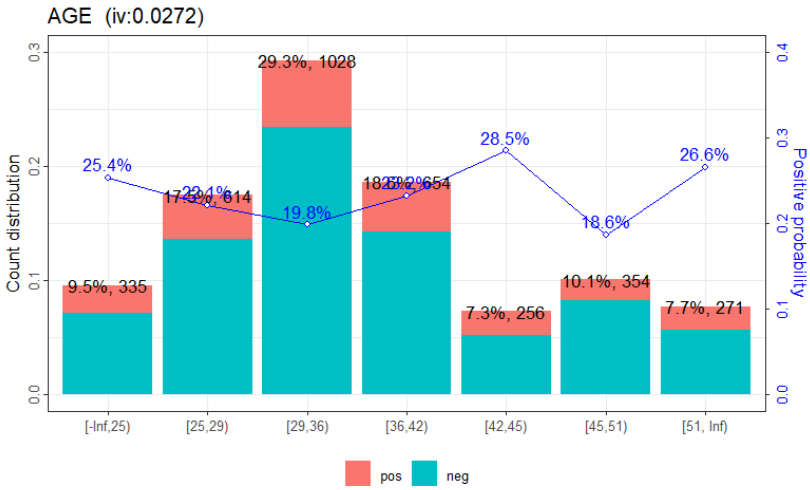
Binning các biến và trưc quan hóa với biểu đồ WOE ta quan sát một số biểu đồ tiêu biểu như sau:




l

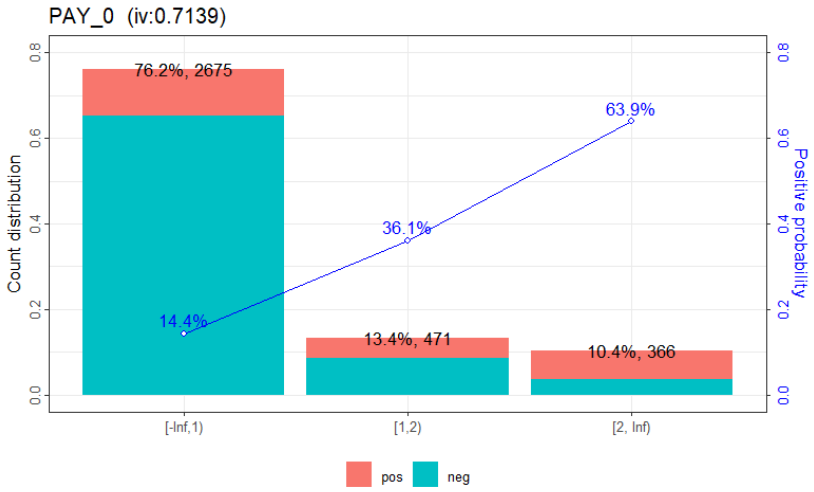
*Hình 2: Biểu đồ WOE của biến LIMIT_BAL*

Nhóm hạn mức thấp nhất ([-Inf, 80000)) có tỷ lệ khách trả nợ cao nhất (~29.1%), nhóm hạn mức cao nhất ([430000, Inf)) có tỷ lệ trả nợ thấp nhất (~13%). Điều này cho thấy hạn mức tín dụng càng cao, khả năng trả nợ càng thấp, biến này có khả năng phân biệt tốt giữa khách trả nợ và không trả nợ. IV = 0.08 (Information Value) cho thấy biến này có mức ảnh hưởng yếu đến biến mục tiêu.




*Hình 3: Biểu đồ WOE của biến AGE*

Biểu đồ cho thấy một mối quan hệ phức tạp và không rõ ràng giữa tuổi và khả năng vỡ nợ do không có một mối quan hệ tuyến tính hoặc một chiều rõ ràng giữa tuổi và khả năng vỡ nợ trong biểu đồ này. Tỷ lệ vỡ nợ có vẻ biến động lên xuống theo các nhóm tuổi khác nhau. Thêm vào đó, giá trị IV thấp cho thấy AGE là một yếu tố dự đoán yếu cho biến mục tiêu.




*Hình 3: Biểu đồ WOE của biến PAY_0*

Biểu đồ cho thấy một mối quan hệ dương và rất mạnh giữa mức độ chậm trả trong tháng trước và khả năng vỡ nợ trong tháng tiếp theo. Khách hàng càng có lịch sử trả nợ tồi tệ (chậm trả nhiều tháng), thì khả năng họ không trả được nợ trong tháng tới càng cao.

Sau đó, ta hồi quy mô hình Logistic theo các biến WOE đã chọn rồi lọc biến bằng stepwise và được mô hình hồi quy mới tốt hơn gồm các biến LIMIT_BAL_woe, AGE_woe, PAY_0_woe, PAY_3_woe, PAY_5_woe, PAY_AMT1_woe, PAY_AMT2_woe và PAY_AMT5_woe. Kết quả ước lượng cho thấy các biến AGE_woe, PAY_0_woe, PAY_3_woe, PAY_5_woe, PAY_AMT1_woe và PAY_AMT2_woe đều có ý nghĩa thống kê ở mức ý nghĩa 5% (p-value < 0.05), trong khi biến LIMIT_BAL_woe chỉ có ý nghĩa ở mức 10% và PAY_AMT5_woe không có ý nghĩa thống kê (p-value > 0.1).

Hệ số của biến PAY_0_woe là 0.75934, cho thấy khi các yếu tố khác giữ nguyên, người có lịch sử thanh toán kém trong tháng gần nhất (PAY_0) có xác suất không trả nợ đúng hạn cao hơn. Tương tự, các biến liên quan đến số tiền thanh toán (PAY_AMT1_woe, PAY_AMT2_woe) có hệ số dương, cho thấy mối quan hệ thuận chiều giữa các khoản thanh toán này và xác suất vỡ nợ, có thể do các khoản này phản ánh lịch sử nợ cũ thay vì khả năng tài chính hiện tại.

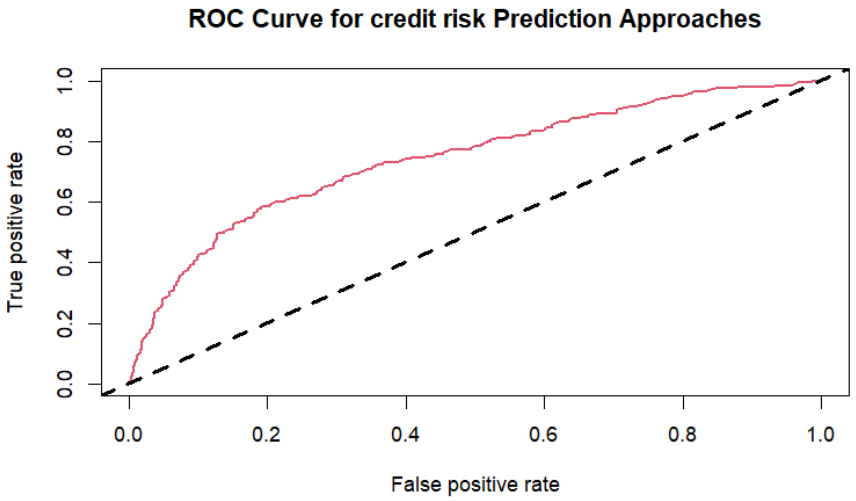
Đánh giá mô hình:

- Confusion Matrix:

| **Dự đoán \ Thực tế** | 0 | 1 |
|---|---|---|
| 0 | 1114 | 237 |
| 1 | 54 | 82 |

- Sensitivity: 60.3%
- Specificity: 82.5%

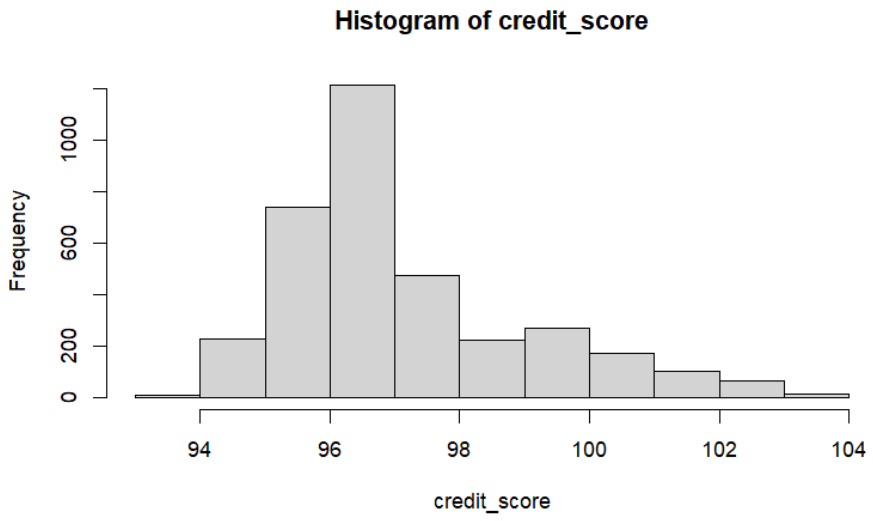
- Accuracy: 80.6%
- AUC: 0.7424, cho thấy mô hình có khả năng phân biệt tốt giữa các lớp.
- Gini: 0.4849, giá trị này càng cao càng cho thấy mô hình có hiệu suất phân loại tốt.
- Đường cong ROC (đường màu đỏ) nằm phía trên đường chéo (đường đứt nét) cho thấy mô hình hoạt động tốt hơn việc phân loại ngẫu nhiên.




*Hình 4: Đường ROC của mô hình Logit theo biến WOE*

Sau đó, ta tính toán credit score thủ công theo công thức dựa trên xác suất dự đoán từ mô hình hồi quy logistic theo biến WOE. Xác suất này phản ánh khả năng khách hàng sẽ vỡ nợ trong tháng tới. Credit score càng cao, càng thể hiện khả năng vỡ nợ càng cao, ngược lại nếu credit score thấp, nghĩa là khả năng vỡ nợ của khách hàng thấp. Cụ thể, kết quả điểm của 10 khách hàng đầu như sau:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 101.08 | 96.12 | 96.69 | 93.91 | 97.49 | 97.3 | 96.2 | 95.2 | 99.4 | 96.11 |




*Hình 5: Đồ thị Histogram của bảng điểm tín dụng*

**Nhận xét:**

Biểu đồ histogram cho thấy điểm tín dụng phân bố chủ yếu trong khoảng từ 94 đến 104, với tần suất cao nhất tại khoảng 96–97, nơi có hơn 1000 khách hàng. Phân bố có xu hướng tập trung, cho thấy phần lớn khách hàng trong tập huấn luyện có đặc điểm tương đối giống nhau về rủi ro tín dụng theo đánh giá của mô hình.

Tuy nhiên, do thang điểm được tính thủ công chưa được đảo chiều và chỉ dao động trong khoảng hẹp, việc sử dụng điểm này để phân loại rủi ro khách hàng trên thực tế còn hạn chế. Khoảng cách giữa nhóm "rủi ro cao" và "rủi ro thấp" chưa rõ ràng, gây khó khăn cho việc ra quyết định tín dụng. Do đó, ta nên dùng lệnh scorecard() trong R để dễ dàng trong việc phân nhóm khách hàng với số điểm dưới 600 là khả năng vỡ nợ cao, 600-700 là khả năng vỡ nợ trung bình và trên 700 là khả năng trả nợ tốt.

### 4.3 Mô hình Random Forest

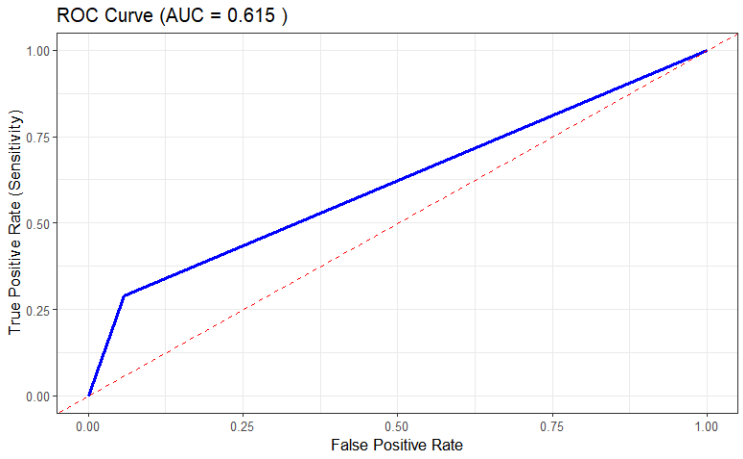
Ta chia dữ liệu thành hai tập (70% train, 30% test) và sau đó huấn luyện mô hình Random Forest với 100 cây quyết định. Kết quả huấn luyện cho thấy tỷ lệ lỗi OOB (Out-Of-Bag error rate) đạt khoảng 20.64%, tương đương độ chính xác trên tập huấn luyện khoảng 79.36%.

Đánh giá mô hình:

- Confusion Matrix:

| | Thực tế = 0 | Thực tế = 1 |
|---|---|---|
| Dự đoán = 0 | 1100 | 227 |
| Dự đoán = 1 | 68 | 92 |

- Accuracy (Độ chính xác): 80.16%, mô hình phân loại đúng khoảng 80% quan sát trên tập kiểm tra
- Sensitivity (True Positive Rate – dự đoán đúng lớp 0): 57.5%, mô hình nhận diện trung bình những người sẽ không vỡ nợ
- Specificity (True Negative Rate – dự đoán đúng lớp 1): 82.89%, mô hình hiệu quả trong việc phát hiện người có khả năng vỡ nợ
- Balanced Accuracy: 61.51%
- Kappa: 0.2811 (mức độ đồng thuận thấp)
- AUC (Area Under Curve): 0.684 cho thấy mô hình có khả năng phân loại ở mức khá.
- Đường ROC nằm phía trên đường đỏ chứng tỏ mô hình có khả năng phân biệt giữa hai lớp tốt hơn đoán ngẫu nhiên.




*Hình 6: Đường ROC của mô hình Random Forest*

### 4.4 Mô hình K-Nearest Neighbors (KNN)

Để đánh giá khả năng phân loại của mô hình KNN, ta thực hiện chuẩn hóa dữ liệu đầu vào và thử nghiệm với nhiều giá trị k từ 1 đến 20. Sau khi vẽ đồ thị biểu diễn độ chính xác (accuracy) của mô hình đối với từng giá trị k, ta chọn giá trị k = 16 là tối ưu, với độ chính xác đạt 80.12% trên tập kiểm tra.

Kết quả phân loại trên tập kiểm tra như sau:

- Confusion Matrix:

| | Thực tế = 0 | Thực tế = 1 |
|---|---|---|
| Dự đoán = 0 | 1106 | 237 |
| Dự đoán = 1 | 61 | 95 |

- Độ chính xác (Accuracy): 80.12%, với khoảng tin cậy 95% từ 78.01% đến 82.11%.

- Sensitivity (True Positive Rate): 61.89%, cho thấy mô hình có khả năng nhận diện trung bình các khách hàng có nguy cơ vỡ nợ.
- Specificity (True Negative Rate): 82.35%, chỉ ra rằng mô hình phân biệt tốt các khách hàng không có nguy cơ vỡ nợ.
- Kappa: 0.2886, cho thấy mô hình có mức độ đồng thuận khá thấp so với dự đoán ngẫu nhiên.
- Balanced Accuracy: 61.69%, phản ánh sự cân bằng giữa khả năng phát hiện cả hai lớp (người có nguy cơ vỡ nợ và không có).

# 5 KẾT LUẬN

Mô hình Logistic Regression (Logit) có hiệu suất tốt nhất về Accuracy, Sensitivity, và Specificity, với khả năng phân biệt khá tốt giữa các lớp. Mặc dù AUC không phải là cao nhất, nhưng mô hình này vẫn hoạt động ổn định và có độ chính xác tốt.

Mô hình Logistic Regression với WOE có khả năng phân biệt tốt nhất về AUC và Gini, nhưng Sensitivity thấp hơn, dẫn đến việc bỏ sót nhiều trường hợp mắc nợ.

Mô hình Random Forest có Specificity cao nhất nhưng lại có Sensitivity thấp, khiến mô hình không phát hiện tốt các trường hợp vỡ nợ.

Mô hình KNN có độ chính xác ổn định và tương đối tốt trong việc nhận diện người không mắc nợ, nhưng không mạnh mẽ trong việc phát hiện người có nguy cơ vỡ nợ như các mô hình khác.

Tóm lại, nếu mục tiêu là phát hiện người vỡ nợ, mô hình Logistic Regression (Logit) có vẻ là lựa chọn tốt nhất. Tuy nhiên, nếu cần một mô hình phân biệt rõ ràng hơn giữa các lớp, mô hình Logistic Regression với WOE có thể mang lại kết quả tốt hơn, mặc dù cần cải thiện Sensitivity.

# TÀI LIỆU THAM KHẢO

[1] *The Global Association of Risk Professionals (2020), Financial Risk Management-FRM, Part 2.*

[2] *Alexander J.McNeil, R.Frey and P.Embrechts (2005), Quantitative Risk Management:Concepts, Techniques and Tools, Princeton University Press.*

## Nguồn dữ liệu:

https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/data