ĐẠI HỌC KINH TẾ QUỐC DÂN

KHOA TOÁN KINH TẾ

Quản trị rủi ro định lượng 2

Bài tập cá nhân

GV hướng dẫn:TS.Nguyễn Thị Liên

Sinh viên thực hiện: Trần Thị Kiều Oanh

Mã sinh viên: 11225075

Lớp : Toán Kinh Tế 64

Hệ đào tạo: Chính quy

Hà Nội, 2025

Mục lục

DANH MỤC BẢNG BIỂU **iii**

DANH MỤC HÌNH **iv**

0.1 GIỚI THIỆU 1

0.1.1 Lý do nghiên cứu 1

0.1.2 Mục tiêu nghiên cứu 1

0.1.3 Tầm quan trọng của nghiên cứu 1

0.1.4 Phạm vi nghiên cứu 1

0.2 CƠ SỞ LÝ THUYẾT 2

0.2.1 Tổng quan về lý thuyết 2

0.2.2 Phương pháp học tập 2

0.2.3 Các mô hình 2

0.3 DỮ LIỆU 3

0.3.1 Mô tả dữ liệu 3

0.3.2 Xử lý Dữ liệu 3

0.3.3 Đánh giá và Lọc Biến Dựa trên Information Value 4

0.4 Xây dựng mô hình hồi quy logistic 4

0.4.1 Giới thiệu về mô hình Logit 4

0.4.2 Xây dựng mô hình 5

0.4.3 Đánh giá mô hình 6

0.5 Random Forest và KNN 6

0.5.1 Giới thiệu về Random Forest và KNN trong phân tích rủi ro tín dụng . . 6

0.5.2 Xây dựng mô hình Random Forest 8
0.5.3 Xây dựng mô hình Cây quyết định 8
0.6 So sánh và xây dựng hệ thống tính điểm 11
0.6.1 So sánh 3 mô hình 11
0.6.2 Xây dựng hệ thống tính điểm tín dụng (Credit Scoring) từ mô hình Logit với WOE 11
0.7 Kết luận và thảo luận 12

Danh sách bảng

1 Mô tả các biến trong bộ dữ liệu Credit Risk 3

2 Chỉ số IV của một số biến 4

3 Các biến có ý nghĩa thống kê trong mô hình Logistic Regression 5

4 Ma trận nhầm lẫn trên tập kiểm tra 6

5 So sánh hiệu suất các mô hình 11

Danh sách hình vẽ

1 Biểu đồ ROC - RandomForest 8

2 Đồ thị cây quyết định 9

3 Biểu đồ ROC - KNN 10

4 Biểu đồ cắt tỉa cây quyết định 10

5 Biểu đồ phân phối điểm tín dụng 12

0.1 GIỚI THIỆU

0.1.1 Lý do nghiên cứu

Trong lĩnh vực tài chính – ngân hàng, việc đánh giá khả năng trả nợ của khách hàng là một yếu tố then chốt nhằm giảm thiểu rủi ro tín dụng và tối ưu hóa hoạt động cho vay. Tuy nhiên, với sự gia tăng nhanh chóng của số lượng hồ sơ tín dụng, việc thẩm định theo phương pháp thủ công trở nên không hiệu quả và thiếu chính xác. Do đó, nghiên cứu ứng dụng các mô hình dự báo như Logistic Regression kết hợp xử lý dữ liệu thông minh (WOE, IV) ngày càng trở nên cần thiết nhằm hỗ trợ ra quyết định cho các tổ chức tài chính.

0.1.2 Mục tiêu nghiên cứu

Mục tiêu của nghiên cứu này là xây dựng một mô hình Logistic Regression nhằm dự báo khả năng vỡ nợ của khách hàng dựa trên các đặc điểm nhân khẩu học và lịch sử tín dụng. Thông qua đó, đánh giá hiệu quả của mô hình trên tập dữ liệu thực tế bằng các chỉ tiêu thống kê như Accuracy, AUC, GINI, từ đó tạo nền tảng cho việc xây dựng hệ thống chấm điểm tín dụng (credit scoring).

0.1.3 Tầm quan trọng của nghiên cứu

Việc phát triển một hệ thống dự báo rủi ro tín dụng hiệu quả không chỉ giúp các tổ chức tín dụng kiểm soát rủi ro mà còn góp phần tối ưu hóa lợi nhuận bằng cách nhận diện chính xác đối tượng khách hàng tiềm năng. Ngoài ra, kết quả nghiên cứu cũng đóng góp cho hoạt động phân tích dữ liệu lớn (Big Data Analytics) trong lĩnh vực tài chính, thể hiện tính ứng dụng cao của các phương pháp phân tích định lượng trong thực tiễn.

0.1.4 Phạm vi nghiên cứu

Nghiên cứu tập trung vào việc xây dựng mô hình Logistic Regression nhằm dự đoán rủi ro tín dụng của khách hàng. Các kỹ thuật tiền xử lý dữ liệu như biến đổi Weight of Evidence (WOE) và chọn lọc biến bằng chỉ số Information Value (IV) được áp dụng để cải thiện chất lượng dữ liệu đầu vào. Nghiên cứu giới hạn trong việc khai thác các phương pháp hồi quy tuyến tính cho bài toán phân loại nhị phân và không mở rộng sang các mô hình phức tạp khác như cây quyết định, Random Forest hoặc XGBoost. Việc đánh giá mô hình được thực hiện dựa trên các tiêu chí như Accuracy, AUC, và GINI nhằm phản ánh khả năng phân biệt và hiệu quả dự báo

của mô hình.

0.2 CƠ SỞ LÝ THUYẾT

0.2.1 Tổng quan về lý thuyết

Logistic Regression là một phương pháp thống kê phổ biến dùng để dự báo xác suất của một biến phụ thuộc nhị phân dựa trên một hoặc nhiều biến độc lập. Trong bối cảnh tín dụng, Logistic Regression được sử dụng để dự đoán khả năng khách hàng sẽ vỡ nợ hay không. Khả năng tuyến tính hóa mối quan hệ giữa các biến đầu vào và xác suất xảy ra sự kiện, cùng với việc giải thích dễ dàng thông qua odds ratio, khiến Logistic Regression trở thành lựa chọn phù hợp trong lĩnh vực quản trị rủi ro tín dụng.

0.2.2 Phương pháp học tập

Phương pháp học tập sử dụng trong nghiên cứu là học có giám sát (supervised learning), cụ thể với bài toán phân loại nhị phân (binary classification). Mô hình Logistic Regression được huấn luyện trên tập dữ liệu mà nhãn đã được biết trước. Quy trình học tập bao gồm: biến đổi dữ liệu bằng WOE để chuẩn hóa các biến đầu vào, lựa chọn đặc trưng quan trọng bằng IV và Backward Stepwise Selection, sau đó huấn luyện mô hình Logistic Regression trên tập train và đánh giá hiệu quả dự báo trên tập test. Phương pháp tối ưu hóa mô hình dựa trên việc loại bỏ dần các biến không có ý nghĩa thống kê nhằm đạt được mô hình đơn giản nhưng hiệu quả cao.

0.2.3 Các mô hình

Trong khuôn khổ nghiên cứu này, Logistic Regression là mô hình chủ đạo được sử dụng. Ngoài ra, để tăng tính ổn định và khả năng giải thích của mô hình, kỹ thuật Weight of Evidence (WOE) được áp dụng nhằm biến đổi các biến đầu vào thành dạng số hóa phù hợp với mối quan hệ tuyến tính trong mô hình Logistic. Để lựa chọn các biến đầu vào tối ưu, phương pháp Information Value (IV) và Stepwise Backward Selection được sử dụng nhằm loại bỏ các biến không có ý nghĩa thống kê hoặc ít đóng góp cho mô hình.

0.3 DỮ LIỆU

0.3.1 Mô tả dữ liệu

Bộ dữ liệu Credit Risk Dataset được lấy trên tràn kaggle chứa thông tin về các khoản vay và đặc điểm của người vay, với mục tiêu phân tích và phân loại nguy cơ vỡ nợ. Bộ dữ liệu gồm 12 đặc trưng, bao gồm thông tin về tuổi, thu nhập, tình trạng sở hữu nhà, mục đích vay, xếp hạng tín dụng, số tiền vay, lãi suất vay và tỷ lệ vỡ nợ của các khoản vay. Các đặc trưng chính bao gồm:

Biến	Ý nghĩa	Ghi chú
person_age	Tuổi của người vay	Phổ biến từ 20 - 82 tuổi
person_income	Thu nhập hàng năm	Từ 4,000 USD đến hơn 6 triệu USD (đa số dưới 300,000 USD)
person_home_ownership	Hình thức sở hữu nhà	Thuê (RENT - 50%), Thế chấp (MORTGAGE - 41%), Khác (8%)
person_emp_length	Thời gian làm việc (năm)	0 - 10+ năm
loan_intent	Mục đích vay	Các nhóm phổ biến: Giáo dục, Y tế, Cá nhân, Hợp nhất nợ,...
loan_grade	Xếp hạng tín dụng khoản vay	A, B, C,... (A và B chiếm hơn 60%)
loan_amnt	Số tiền vay	Từ 500 USD đến 35,000 USD
loan_int_rate	Lãi suất vay (%)	Từ 5.42% đến 23.2%
loan_status	Tình trạng khoản vay	0: Không vỡ nợ, 1: Vỡ nợ
loan_percent_income	Tỷ lệ khoản vay trên thu nhập	Thường dưới 0.5
cb_person_default_on_file	Từng vỡ nợ trong quá khứ (có/không)	Phần lớn là không
cb_preson_cred_hist_length	Độ dài lịch sử tín dụng (năm)	0 - 22 năm

Bảng 1: Mô tả các biến trong bộ dữ liệu Credit Risk

0.3.2 Xử lý Dữ liệu

Trước khi thực hiện các phân tích tiếp theo, bộ dữ liệu đã được xử lý để đảm bảo tính đầy đủ và chính xác. Dữ liệu bao gồm nhiều biến, trong đó một số biến có giá trị thiếu (missing values). Các bước xử lý dữ liệu bao gồm việc kiểm tra các giá trị thiếu, xử lý các giá trị này,để có cái nhìn tổng quan về dữ liệu.

Phân Tích Trạng Thái Khoản Vay (`loan_status`)

Biến `loan_status` cho biết tình trạng của khoản vay, với giá trị 0 đại diện cho khoản vay không bị vỡ nợ và 1 đại diện cho khoản vay bị vỡ nợ. Dựa trên bảng phân phối, ta thấy rằng 78.18% các khoản vay không bị vỡ nợ (giá trị 0), trong khi 21.82% các khoản vay bị vỡ nợ (giá trị 1).

Xử Lý Dữ Liệu Thiếu (Missing Data)

Sau khi kiểm tra,phát hiện rằng một số biến có giá trị thiếu, chẳng hạn như `person_emp_length` (thời gian làm việc) và `loan_int_rate` (lãi suất vay). Để xử lý,điền các giá trị thiếu bằng

giá trị trung vị của các biến tương ứng. Cụ thể, giá trị trung vị của `person_emp_length` là 4 năm, và giá trị trung vị của `loan_int_rate` là 10.99%. Sau khi điền giá trị thiếu, bộ dữ liệu được hoàn chỉnh và sẵn sàng cho các phân tích tiếp theo.

0.3.3 Đánh giá và Lọc Biến Dựa trên Information Value

Để nâng cao chất lượng dự báo rủi ro tín dụng,tiến hành đánh giá mức độ đóng góp thông tin của các biến độc lập thông qua chỉ số **Information Value (IV)**. Cụ thể,sử dụng hàm `create_infotables` từ thư viện `Information` để tính toán IV cho từng biến dựa trên biến mục tiêu `loan_status`.

Tên biến	IV	Mức độ ảnh hưởng
loan_percent_income	0.887	Rất mạnh
loan_grade	0.882	Rất mạnh
loan_int_rate	0.637	Mạnh
person_income	0.487	Trung bình
cb_person_default_on_file	0.164	Yếu

Bảng 2: Chỉ số IV của một số biến

Kết quả cho thấy, hai biến `person_age` (tuổi của người vay) và `cb_person_cred_hist_length` (độ dài lịch sử tín dụng) có giá trị IV nhỏ hơn 0.02, cho thấy mức độ dự báo yếu đối với khả năng vỡ nợ. Theo nguyên tắc thực hành trong phân tích tín dụng, các biến có IV thấp hơn ngưỡng này sẽ được loại bỏ nhằm tránh gây nhiễu cho mô hình.

Do đó,loại bỏ hai biến này khỏi bộ dữ liệu, chỉ giữ lại những biến có giá trị thông tin cao hơn để phục vụ cho các bước phân tích và xây dựng mô hình tiếp theo.

0.4 Xây dựng mô hình hồi quy logistic

0.4.1 Giới thiệu về mô hình Logit

Mô hình Logit, hay còn gọi là mô hình hồi quy logistic, là một phương pháp phân tích thống kê được sử dụng phổ biến để dự đoán xác suất xảy ra của một biến nhị phân (binary outcome), tức là biến phụ thuộc chỉ nhận một trong hai giá trị, ví dụ như "vỡ nợ" hoặc "không vỡ nợ". Khác với mô hình hồi quy tuyến tính thông thường vốn giả định mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc liên tục, mô hình Logit áp dụng hàm logistic để chuyển đổi đầu ra thành một xác suất nằm trong khoảng từ 0 đến 1.

Cụ thể, mô hình Logit mô tả mối quan hệ giữa các biến độc lập $X_1, X_2, ..., X_k$ và xác suất xảy ra sự kiện (ví dụ: khách hàng vỡ nợ) thông qua hàm:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

Trong đó:

- $P(Y = 1|X)$ là xác suất mà sự kiện xảy ra (Y = 1),
- β_0 là hệ số chặn (intercept),
- $\beta_1, ..., \beta_k$ là các hệ số hồi quy tương ứng với các biến độc lập $X_1, ..., X_k$.

Một ưu điểm nổi bật của mô hình Logit là khả năng giải thích kết quả dưới dạng xác suất, đồng thời giúp đánh giá mức độ ảnh hưởng của từng biến giải thích đối với khả năng xảy ra của sự kiện. Trong lĩnh vực tài chính – tín dụng, mô hình Logit được sử dụng rộng rãi để xây dựng hệ thống chấm điểm tín dụng (credit scoring), dự báo khả năng vỡ nợ của khách hàng, và hỗ trợ ra quyết định trong quản trị rủi ro.

0.4.2 Xây dựng mô hình

Chia dữ liệu thành hai tập: 70% dùng để huấn luyện mô hình và 30% để kiểm định. Trên tập huấn luyện, các biến được biến đổi theo phương pháp WOE (Weight of Evidence), giúp cải thiện khả năng phân loại và tăng cường ý nghĩa diễn giải cho mô hình logistic. Mô hình hồi quy logistic ban đầu bao gồm toàn bộ các biến còn lại. Tuy nhiên, để đơn giản hóa và tối ưu hiệu quả,sử dụng phương pháp lựa chọn biến stepwise để loại bỏ những biến không có ý nghĩa thống kê. Mô hình cuối cùng bao gồm 7 biến đầu vào có ảnh hưởng đáng kể, trong đó đáng chú ý là `loan_intent`, `loan_grade`, và `loan_percent_income` với hệ số ước lượng lớn và giá trị p rất nhỏ (p < 2e-16).

Biến	**Hệ số (Estimate)**	**p-value**	**Ý nghĩa thống kê**
`person_income_woe`	0.79	$< 2 \times 10^{-16}$	Có
`person_home_ownership_woe`	0.97	$< 2 \times 10^{-16}$	Có
`person_emp_length_woe`	0.27	0.0021	Có
`loan_intent_woe`	1.50	$< 2 \times 10^{-16}$	Có
`loan_grade_woe`	1.09	$< 2 \times 10^{-16}$	Có
`loan_int_rate_woe`	0.19	1.4×10^{-5}	Có
`loan_percent_income_woe`	1.08	$< 2 \times 10^{-16}$	Có

Bảng 3: Các biến có ý nghĩa thống kê trong mô hình Logistic Regression

	Dự đoán 0	Dự đoán 1
Thực tế 0	7297	354
Thực tế 1	832	1257

Bảng 4: Ma trận nhầm lẫn trên tập kiểm tra

0.4.3 Đánh giá mô hình

Để đánh giá mô hình,kiểm định trên cả tập huấn luyện và tập kiểm tra. Kết quả cho thấy, mô hình đạt được độ chính xác cao với tỷ lệ dự đoán đúng trên tập test là 87,82%. Ngoài ra, chỉ số AUROC là 0.882 và hệ số GINI là 0.764, cho thấy mô hình có khả năng phân biệt mạnh giữa khách hàng có và không có nguy cơ vỡ nợ. Thêm vào đó, độ nhạy đạt 95,37% và độ đặc hiệu đạt 60,17%, phản ánh sự cân bằng giữa khả năng phát hiện đúng và tránh nhầm lẫn.

0.5 Random Forest và KNN

0.5.1 Giới thiệu về Random Forest và KNN trong phân tích rủi ro tín dụng

Trong lĩnh vực phân tích rủi ro tín dụng, bên cạnh các mô hình thống kê truyền thống như hồi quy logistic, các phương pháp học máy hiện đại như **Random Forest** và **K-Nearest Neighbors (KNN)** đã chứng minh được hiệu quả vượt trội trong việc dự đoán khả năng vỡ nợ của khách hàng. Nhờ khả năng học từ dữ liệu phức tạp và không yêu cầu giả định phân phối, hai phương pháp này ngày càng được ứng dụng rộng rãi trong các hệ thống chấm điểm tín dụng và quản trị rủi ro.

Phương pháp Random Forest

Random Forest là một thuật toán học máy thuộc nhóm mô hình *ensemble* (tập hợp), được xây dựng dựa trên nhiều cây quyết định (*decision trees*) hoạt động song song. Cụ thể, mỗi cây trong rừng được huấn luyện trên một tập con ngẫu nhiên của dữ liệu, và chỉ sử dụng một tập hợp con của các biến đầu vào. Nhờ đó, phương pháp này giúp giảm thiểu độ lệch (bias) và độ phương sai (variance) của mô hình. Kết quả cuối cùng được xác định bằng cách lấy biểu quyết số đông từ tất cả các cây đối với bài toán phân loại.

Ưu điểm lớn của Random Forest trong phân tích rủi ro tín dụng bao gồm khả năng xử lý dữ liệu có nhiều biến đầu vào phức tạp, khả năng chống quá khớp (*overfitting*) tốt hơn so với cây quyết định đơn lẻ, và khả năng đánh giá tầm quan trọng của từng biến đối với xác suất vỡ nợ. Ngoài ra, mô hình này cũng dễ dàng mở rộng và tích hợp vào hệ thống tính điểm tín dụng

hiện có. Tuy nhiên, vì Random Forest là một mô hình dạng "hộp đen", nên việc diễn giải và xây dựng thang điểm tín dụng trực tiếp từ mô hình này thường gặp nhiều thách thức hơn so với mô hình Logit.

Phương pháp K-Nearest Neighbors (KNN)

Trái ngược với Random Forest, KNN là một thuật toán đơn giản nhưng lại rất hiệu quả trong nhiều tình huống. Phương pháp này dựa trên nguyên tắc tính khoảng cách giữa điểm dữ liệu cần dự đoán và các điểm đã biết trong không gian đặc trưng. Khi có một quan sát mới, KNN sẽ tìm ra K quan sát gần nhất (dựa trên khoảng cách Euclid hoặc Manhattan), và phân loại dựa vào tỉ lệ nhãn của các "hàng xóm" này.

Trong ngữ cảnh đánh giá rủi ro tín dụng, KNN có thể phát hiện các mẫu khách hàng tương đồng dựa trên hồ sơ tín dụng, và nhờ đó phân loại họ vào các nhóm rủi ro tương ứng. Hơn nữa, phương pháp này không yêu cầu giả định về phân phối dữ liệu, rất thích hợp trong các tập dữ liệu thực tế. Tuy vậy, KNN thường yêu cầu chuẩn hóa dữ liệu trước khi sử dụng, do sự phụ thuộc mạnh vào khoảng cách. Bên cạnh đó, hiệu suất của mô hình sẽ giảm rõ rệt nếu số lượng biến (chiều) quá lớn – một hiện tượng được biết đến là "lời nguyền của chiều".

Ứng dụng vào xây dựng thang điểm tín dụng

Mặc dù cả Random Forest và KNN đều không trực tiếp sinh ra hệ số như hồi quy logistic (vốn thuận lợi cho việc xây dựng bảng điểm), nhưng chúng vẫn có thể được sử dụng hiệu quả trong việc ước lượng xác suất vỡ nợ. Từ xác suất đó, có thể ánh xạ sang thang điểm tín dụng (*credit score*) thông qua các công thức tuyến tính hoặc logistic.

Một công thức thường được sử dụng là:

$$\text{Credit Score} = \text{Offset} + \text{Factor} \times \log\left(\frac{1-p}{p}\right)$$

Trong đó, p là xác suất vỡ nợ được ước lượng từ mô hình Random Forest hoặc KNN. Offset và Factor có thể được điều chỉnh tùy theo thang điểm mong muốn (ví dụ: quy đổi về thang điểm từ 300 đến 850). Thang điểm này giúp các tổ chức tín dụng dễ dàng phân loại khách hàng thành các nhóm rủi ro khác nhau, từ đó đưa ra các quyết định cho vay phù hợp và hiệu quả hơn.

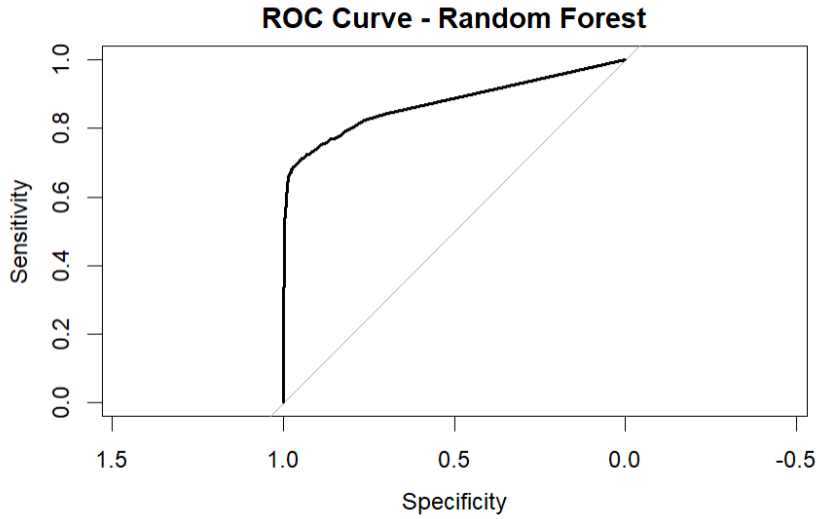
0.5.2 Xây dựng mô hình Random Forest

Đầu tiên, mô hình Random Forest được xây dựng bằng cách sử dụng hàm `randomForest()` với các tham số `ntree = 100`, `mtry` = $\sqrt{p}$, và `importance = TRUE` nhằm đánh giá tầm quan trọng của các biến đầu vào. Mô hình được huấn luyện trên tập dữ liệu đã xử lý bằng WOE, sau đó được sử dụng để dự đoán cả nhãn phân loại (0/1) và xác suất vỡ nợ trên tập kiểm tra.

Kết quả đánh giá trên tập kiểm tra thể hiện trong ma trận nhầm lẫn dưới đây:

Predicted / Actual	**0**	**1**
0	7530	713
1	133	1409

Dựa trên kết quả này, độ chính xác (Accuracy) đạt **91.35%**, Precision đạt **66.40%**, Recall là **91.37%**, và F1-score đạt **76.91%**. Đặc biệt, diện tích dưới đường cong ROC (AUC) lên tới **0.8737**, phản ánh khả năng phân loại tốt giữa hai nhóm khách hàng.



Hình 1: Biểu đồ ROC - RandomForest

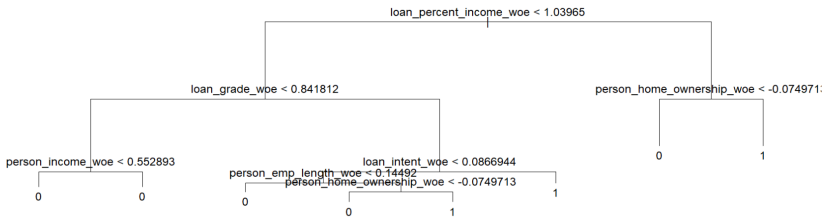
Biểu đồ ROC xác nhận rằng mô hình Random Forest có hiệu năng phân biệt rõ rệt giữa nhóm vỡ nợ và không vỡ nợ.

0.5.3 Xây dựng mô hình Cây quyết định

Tiếp theo, mô hình Cây quyết định được xây dựng sử dụng thư viện `tree`. Các tham số được thiết lập bao gồm: `mincut = 10`, `minsize = 20`, và `mindev = 0.01` nhằm kiểm

soát độ sâu và tính đơn giản của cây. Kết quả cho thấy cây có **8 nút lá**, và các biến được lựa chọn bao gồm:

- `loan_percent_income_woe`
- `loan_grade_woe`
- `person_income_woe`
- `loan_intent_woe`
- `person_emp_length_woe`
- `person_home_ownership_woe`

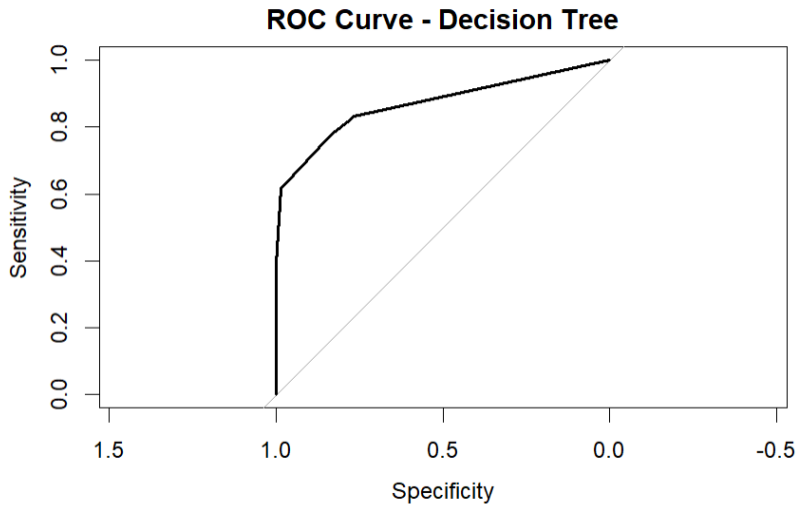


Hình 2: Đồ thị cây quyết định

Ma trận nhầm lẫn của mô hình trên tập kiểm tra như sau:

Predicted / Actual	**0**	**1**
0	7548	815
1	115	1307

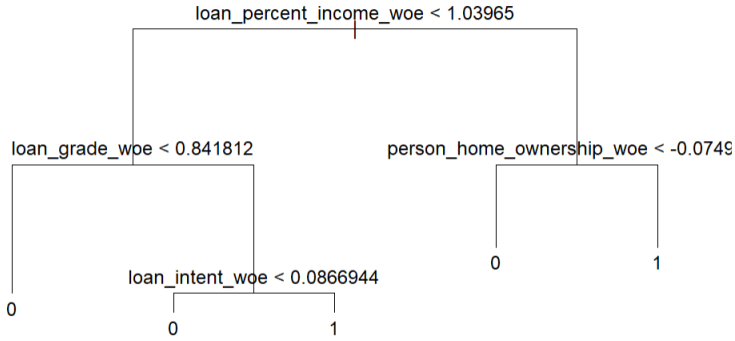
Từ đó, ta tính được các chỉ số đánh giá: độ chính xác là **90.50%**, Precision là **61.59%**, Recall đạt **91.91%** và F1-score là **73.66%**. Đồng thời, chỉ số AUC cũng đạt mức cao là **0.8696**, thể hiện mô hình có hiệu quả phân loại tốt, mặc dù thấp hơn một chút so với Random Forest.



Hình 3: Biểu đồ ROC - KNN

Cắt tỉa cây quyết định (Pruning)

Để cải thiện tính đơn giản và khả năng tổng quát của cây quyết định, nhóm đã thực hiện cắt tỉa cây bằng cách sử dụng hàm `cv.tree()` nhằm lựa chọn số nút lá tối ưu.



Hình 4: Biểu đồ cắt tỉa cây quyết định

Kết quả cho thấy cây tốt nhất có **5 nút lá**, giúp mô hình trở nên đơn giản hơn nhưng vẫn giữ được hiệu quả phân loại ổn định. Việc cắt tỉa cây giúp tăng tính dễ hiểu của mô hình, phù hợp với yêu cầu minh bạch trong lĩnh vực tín dụng.

0.6 So sánh và xây dựng hệ thống tính điểm

0.6.1 So sánh 3 mô hình

Mô hình	Accuracy	AUC	Gini
Logistic với WOE	87.82%	0.882	0.764
KNN	87.40%	0.844	0.689
Random Forest	91.35%	0.874	0.748

Bảng 5: So sánh hiệu suất các mô hình

Dựa trên bảng so sánh trên, có thể thấy rằng cả ba mô hình Logistic với WOE, KNN và Random Forest đều mang lại kết quả dự đoán tốt trong bài toán phân loại rủi ro tín dụng. Trong đó, mô hình Random Forest đạt độ chính xác cao nhất với **91.35%**, đồng thời đạt AUC là **0.8737** và hệ số GINI là **0.748**, thể hiện khả năng phân biệt rõ ràng giữa khách hàng vỡ nợ và không vỡ nợ. Trong khi đó, mô hình Logistic với WOE đạt AUC cao nhất là **0.882** và GINI là **0.764**, cho thấy hiệu suất phân biệt cũng rất tốt, đặc biệt trong việc phát hiện đúng khách hàng vỡ nợ với độ nhạy cao (95.37%).

Tuy nhiên, mô hình KNN có hiệu năng thấp hơn so với hai mô hình còn lại, với AUC là **0.844** và GINI là **0.689**. Điều này cho thấy khả năng phân loại của KNN tuy khá ổn nhưng chưa đạt mức tối ưu trong bài toán này.

Tóm lại, giữa ba mô hình, Random Forest có hiệu suất tổng thể tốt nhất, trong khi Logit với WOE là lựa chọn hiệu quả, dễ triển khai hơn do tính đơn giản và mô hình KNN tỏ ra không phù hợp trong bối cảnh này. Trong trường hợp ưu tiên được đặt vào việc phát hiện các trường hợp có nguy cơ vỡ nợ, mô hình Logit với WOE là mô hình được chọn

0.6.2 Xây dựng hệ thống tính điểm tín dụng (Credit Scoring) từ mô hình Logit với WOE

Kết quả từ mô hình hồi quy logistic với WOE được sử dụng làm cơ sở để xây dựng bảng điểm tín dụng (scorecard). Mục đích của việc này là chuyển đổi xác suất vỡ nợ thành điểm tín dụng, giúp dễ dàng đánh giá và phân loại khách hàng dựa trên mức độ rủi ro tín dụng của họ.

Để tính điểm tín dụng, trước tiên ta tính **z-score** – một đại lượng biểu thị tỷ lệ giữa xác suất vỡ nợ và xác suất không vỡ nợ – theo công thức:

$$z\_score = \log\left(\frac{p}{1-p}\right)$$

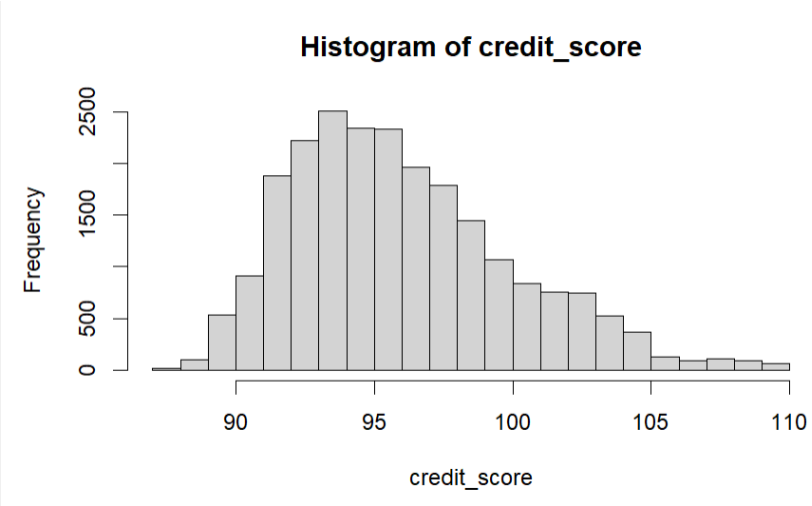
Trong đó, p là xác suất vỡ nợ của khách hàng được dự báo bởi mô hình.

Sau khi tính được z-score, điểm tín dụng được xác định theo công thức:

$$\text{credit\_score} = 100 + 2 \times z\_score$$

Ở đây, **100** là giá trị *offset*, và **2** là *scaling factor*. *Scaling factor* có tác dụng điều chỉnh tốc độ thay đổi điểm số khi xác suất vỡ nợ thay đổi, trong khi *offset* là mức điểm cơ sở ban đầu.

Cuối cùng, điểm tín dụng được tính toán cho từng khách hàng sẽ được sử dụng để phân loại khách hàng theo mức độ rủi ro tín dụng của họ. Biểu đồ phân phối điểm tín dụng cũng được xây dựng nhằm trực quan hóa kết quả và hỗ trợ phân tích mức độ rủi ro của các khách hàng trong tập huấn luyện.



Hình 5: Biểu đồ phân phối điểm tín dụng

Phân bố điểm số cho thấy phần lớn khách hàng nằm trong khoảng 97 đến 104 điểm, phản ánh phân phối hợp lý giữa các mức rủi ro tín dụng.

0.7 Kết luận và thảo luận

Nghiên cứu này đã tiến hành so sánh hiệu suất của các mô hình phân loại bao gồm:mô hình Logit với biến đã chuyển đổi theo Weight of Evidence (WOE), RandomForest và K-Nearest Neighbors (KNN) trong việc dự đoán khả năng vỡ nợ của khách hàng. Kết quả cho thấy mô hình Logit với WOE thể hiện Sensitivity cao nhất, đồng thời có chỉ số AUC và Gini ổn định, cho thấy khả năng phân biệt tốt giữa hai nhóm khách hàng. Mô hình Random Forest tuy có độ chính xác tổng thể cao nhưng lại không cải thiện đáng kể về Sensitivity, trong khi mô hình KNN thể hiện hiệu suất kém rõ rệt do Sensitivity rất thấp.

Từ đó, mô hình Logit với WOE được lựa chọn để xây dựng thang điểm tín dụng (scorecard). Điểm tín dụng được tính toán từ xác suất vỡ nợ thông qua điểm z-score và được quy đổi về thang điểm tuyến tính để dễ áp dụng trong thực tiễn. Phân phối điểm tín dụng cho thấy phần lớn khách hàng nằm trong khoảng điểm từ 97 đến 104, cho thấy đây là nhóm chiếm tỷ trọng lớn nhất trong tập dữ liệu huấn luyện.

Kết quả này cho thấy việc áp dụng WOE không chỉ giúp cải thiện hiệu suất mô hình mà còn hỗ trợ việc xây dựng thang điểm tín dụng một cách hiệu quả và dễ diễn giải. Tuy nhiên, nghiên cứu cũng còn một số hạn chế, như việc chưa tối ưu ngưỡng phân loại hay chưa đánh giá mô hình trên dữ liệu ngoài mẫu (out-of-time). Các yếu tố này có thể được xem xét trong các nghiên cứu tiếp theo để nâng cao độ chính xác và tính ứng dụng thực tiễn của mô hình.