BÁO CÁO
Phân loại tín dụng khách hàng theo dữ liệu giao dịch tài chính

Hà Nội, tháng 12 năm 2024

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN NGHIÊN CỨU

1.1. Bối cảnh nghiên cứu

Trong bối cảnh nền kinh tế Việt Nam đang không ngừng phát triển, hoạt động tín dụng ngân hàng giữ vai trò then chốt trong việc thúc đẩy dòng vốn đầu tư và tăng trưởng kinh tế. Tuy nhiên, cùng với sự mở rộng tín dụng, vấn đề **nợ xấu** (Non-Performing Loans - NPLs) đã trở thành mối quan tâm lớn đối với các tổ chức tín dụng và cơ quan quản lý. Nợ xấu không chỉ gây thiệt hại tài chính trực tiếp cho ngân hàng mà còn làm suy yếu lòng tin của khách hàng, ảnh hưởng đến tính ổn định của hệ thống tài chính và nền kinh tế quốc gia.

Tại Việt Nam, Ngân hàng Nhà nước đã ban hành nhiều chính sách để quản lý và kiểm soát nợ xấu. Tuy nhiên, với sự gia tăng của các khoản vay tín chấp, các rủi ro tín dụng tiềm tàng ngày càng trở nên khó kiểm soát. Do đó, việc áp dụng các phương pháp tiên tiến như **phân tích dữ liệu lớn** (Big Data) và **máy học** (Machine Learning) đã được đặt ra như một giải pháp tiềm năng để nâng cao hiệu quả quản lý nợ xấu.

1.2. Mục tiêu nghiên cứu

Nghiên cứu này nhằm đạt được các mục tiêu sau:

(1) **Phân tích tác động của nợ xấu** đối với hệ thống ngân hàng và nền kinh tế, đồng thời làm rõ các yếu tố ảnh hưởng đến nợ xấu.

(2) **Xây dựng các mô hình dự đoán** nợ xấu dựa trên các thuật toán máy học hiện đại như Random Forest, Logistic Regression và Support Vector Machines, từ đó cải thiện độ chính xác và hiệu quả trong việc phân loại khách hàng.

(3) Đề xuất **giải pháp quản lý và giảm thiểu nợ xấu** thông qua các chính sách tín dụng và các công cụ công nghệ.

1.3. Ý nghĩa nghiên cứu

Về lý thuyết: Nghiên cứu đóng góp vào việc phát triển các mô hình dự đoán nợ xấu hiệu quả, kết hợp giữa lý thuyết kinh tế tài chính và công nghệ máy học hiện đại.

Về thực tiễn: Cung cấp cho các ngân hàng một công cụ mạnh mẽ để nhận diện rủi ro tín dụng, từ đó đưa ra các biện pháp xử lý và phòng ngừa nợ xấu kịp thời. Đồng thời, các giải pháp được đề xuất giúp tối ưu hóa quy trình quản lý tín dụng, giảm thiểu tổn thất và gia tăng tính bền vững cho hệ thống tài chính.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Cơ sở lý thuyết về nợ xấu trong ngân hàng

2.1.1. Khái niệm và tầm quan trọng của nợ xấu

Nợ xấu (Non-Performing Loans - NPLs) là các khoản vay mà người vay không thực hiện đầy đủ nghĩa vụ thanh toán nợ gốc và/hoặc lãi trong thời hạn đã thỏa thuận. Theo thông lệ quốc tế, nợ xấu được định nghĩa là các khoản vay quá hạn từ 90 ngày trở lên. Tại Việt Nam, Ngân hàng Nhà nước (NHNN) quy định nợ xấu bao gồm các khoản thuộc nhóm 3, 4 và 5 trong hệ thống phân loại tín dụng.

Việc quản lý nợ xấu là một yếu tố quan trọng đối với sự ổn định và hiệu quả hoạt động của hệ thống ngân hàng. Tỷ lệ nợ xấu cao không chỉ ảnh hưởng đến lợi nhuận của các tổ chức tín dụng mà còn đe dọa đến thanh khoản và khả năng hoạt động của hệ thống tài chính. Đồng thời, nợ xấu làm giảm niềm tin của nhà đầu tư và khách hàng, gây áp lực lớn lên toàn bộ nền kinh tế.

2.1.2. Phân loại nợ xấu theo quy định tại Việt Nam

Theo quy định tại Điều 10 Thông tư 11/2021/TT-NHNN, các tổ chức tín dụng tại Việt Nam thực hiện phân loại nợ thành 5 nhóm, bao gồm nợ đủ tiêu chuẩn (nhóm 1), nợ cần chú ý (nhóm 2), nợ dưới tiêu chuẩn (nhóm 3), nợ nghi ngờ (nhóm 4), và nợ có khả năng mất vốn (nhóm 5). Trong đó, nhóm 3, 4 và 5 được xếp vào nợ xấu.

(1) **Nhóm 1 (Nợ đủ tiêu chuẩn):** Nhóm này bao gồm các khoản nợ trong hạn và được đánh giá là có khả năng thu hồi đầy đủ cả nợ gốc và lãi đúng hạn. Ngoài ra, các khoản nợ quá hạn dưới 10 ngày nhưng vẫn được đánh giá là có khả năng thu hồi đầy đủ cũng thuộc nhóm này.

(2) **Nhóm 2 (Nợ cần chú ý):** Nhóm này bao gồm các khoản nợ quá hạn đến 90 ngày, nhưng chưa thuộc mức độ nghiêm trọng như nhóm nợ xấu. Nhóm này còn bao gồm các khoản nợ được cơ cấu lại kỳ hạn trả nợ lần đầu nhưng vẫn trong hạn thanh toán.

(3) **Nhóm 3 (Nợ dưới tiêu chuẩn):** Nhóm này bao gồm các khoản nợ quá hạn từ 91 đến 180 ngày, hoặc các khoản nợ được gia hạn lần đầu nhưng vẫn chưa có khả năng thanh toán đầy đủ. Ngoài ra, các khoản nợ được miễn hoặc giảm lãi do khách hàng không đủ khả năng trả lãi theo thỏa thuận cũng thuộc nhóm này.

(4) **Nhóm 4 (Nợ nghi ngờ):** Nhóm này bao gồm các khoản nợ quá hạn từ 181 đến 360 ngày. Các khoản nợ cơ cấu lại kỳ hạn trả nợ lần thứ hai hoặc quá hạn trong phạm vi 90 ngày cũng được xếp vào nhóm này. Những khoản nợ này thường có mức độ rủi ro cao và đòi hỏi tổ chức tín dụng phải có các biện pháp xử lý mạnh mẽ.

(5) **Nhóm 5 (Nợ có khả năng mất vốn):** Nhóm này là nhóm nghiêm trọng nhất, bao gồm các khoản nợ quá hạn trên 360 ngày. Ngoài ra, các khoản nợ cơ cấu lại lần thứ ba trở lên hoặc không thể thu hồi sau 60 ngày kể từ khi có quyết định thu hồi cũng thuộc nhóm này. Nợ thuộc nhóm 5 thường không còn khả năng thu hồi, gây tổn thất lớn cho tổ chức tín dụng.

Theo quy định, các khoản nợ thuộc nhóm 3, 4 và 5 được xếp vào nợ xấu. Đây là các khoản nợ có nguy cơ cao, không chỉ gây thiệt hại tài chính mà còn làm giảm uy tín của ngân hàng trên thị trường. Trong đó, nợ thuộc nhóm 5 được xem là nghiêm trọng nhất và thường khiến khách hàng mất khả năng tiếp cận nguồn vốn từ các ngân hàng và tổ chức tín dụng khác.

2.1.3. Tác động của nợ xấu đến hệ thống ngân hàng và nền kinh tế

Nợ xấu gây ảnh hưởng tiêu cực không chỉ đối với tổ chức tín dụng mà còn đối với toàn bộ hệ thống tài chính và nền kinh tế. Đối với ngân hàng, nợ xấu làm tăng chi phí dự phòng rủi ro, giảm lợi nhuận, và hạn chế khả năng cấp tín dụng. Đối với hệ thống tài chính, tỷ lệ nợ xấu cao có thể dẫn đến rủi ro hệ thống, ảnh hưởng tiêu cực đến niềm tin của nhà đầu tư và khách hàng. Về mặt kinh tế, nợ xấu làm giảm dòng vốn đầu tư, kìm hãm tăng trưởng kinh tế và gây áp lực lớn lên các chính sách tài khóa và tiền tệ.

2.2. Cơ sở lý thuyết về thẩm định tín dụng và nguyên tắc 5Cs trong thẩm định tín dụng

2.2.1. Thẩm định tín dụng

Thẩm định tín dụng là quá trình đánh giá và phân tích khả năng tài chính, lịch sử tín dụng, và mức độ rủi ro của khách hàng trước khi quyết định cấp vốn. Theo Altman (1968), đây là một bước quan trọng giúp các tổ chức tài chính giảm thiểu nguy cơ mất vốn thông qua việc sử dụng các chỉ số tài chính và phân tích định lượng. Việc thẩm định không chỉ đảm bảo tính khả thi của dự án hoặc khoản vay mà còn xác định khả năng trả nợ của khách hàng dựa trên dữ liệu thực tế và các công cụ phân tích hiện đại.

2.2.2. Nguyên tắc 5Cs trong thẩm định tín dụng

Nguyên tắc 5Cs là phương pháp đánh giá rủi ro phổ biến, bao gồm 5 yếu tố chính: **Character (Tư cách), Capacity (Năng lực), Capital (Vốn), Collateral (Tài sản đảm bảo), và Conditions (Điều kiện kinh tế)**. Theo Saunders và Cornett (2018), mỗi yếu tố đều đóng vai trò quan trọng trong việc xác định mức độ an toàn khi cấp tín dụng.

(1) **Character (Tư cách):** Character đề cập đến tính cách, đạo đức và sự uy tín của khách hàng trong việc thực hiện các nghĩa vụ tài chính. Lịch sử tín dụng và khả năng quản lý tài chính của khách hàng là các yếu tố quan trọng để đánh giá tư cách. Basel Committee (2006) nhấn mạnh rằng một hồ sơ tín dụng sạch và minh bạch thường là dấu hiệu tích cực, giảm nguy cơ không thực hiện nghĩa vụ thanh toán.

(2) **Capacity (Năng lực):** Capacity đo lường khả năng tài chính của khách hàng trong việc trả nợ thông qua dòng tiền từ hoạt động kinh doanh hoặc thu nhập cá nhân. Altman (1968) đã phát triển các mô hình phân tích tài chính để đánh giá năng lực này, bao gồm việc sử dụng các chỉ số tài chính như tỷ lệ thanh khoản và tỷ lệ nợ/vốn.

(3) **Capital (Vốn):** Capital đại diện cho lượng vốn tự có mà khách hàng đầu tư vào dự án hoặc kế hoạch kinh doanh. Theo Saunders và Cornett (2018), tỷ lệ vốn tự có cao thể hiện cam kết tài chính mạnh mẽ của khách hàng và giảm thiểu rủi ro cho ngân hàng trong trường hợp dự án thất bại.

(4) **Collateral (Tài sản đảm bảo):** Collateral là tài sản mà khách hàng thế chấp để bảo đảm khoản vay. Basel Committee (2006) đã nhấn mạnh rằng giá trị tài sản đảm bảo phải được đánh giá chính xác và có tính thanh khoản cao để ngân hàng có thể thu hồi vốn trong trường hợp xảy ra rủi ro.

(5) **Conditions (Điều kiện kinh tế):** Conditions đề cập đến các yếu tố bên ngoài, bao gồm điều kiện kinh tế vĩ mô và các yếu tố ngành nghề. Saunders và Cornett (2018) cho rằng các biến động kinh tế, như suy thoái kinh tế hoặc lãi suất cao, có thể ảnh hưởng đáng kể đến khả năng trả nợ của khách hàng.

2.3. Cơ sở lý thuyết về mô hình máy học

2.3.1. Logistic Regression

Logistic Regression là một phương pháp thống kê được sử dụng để phân loại dữ liệu, đặc biệt trong các bài toán nhị phân. Mô hình này dự đoán xác suất xảy ra của một sự kiện thông qua hàm sigmoid, giúp chuyển đổi giá trị đầu ra thành một giá trị xác suất trong khoảng từ 0 đến 1 (Hosmer et al., 2013). Logistic Regression thường được áp dụng trong phân loại rủi ro tín dụng, y học và tiếp thị.

2.3.2. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) là một thuật toán phi tham số được sử dụng để phân loại hoặc hồi quy dữ liệu. Nguyên tắc của KNN là xác định "K" điểm gần nhất với điểm dữ liệu mới trong không gian đa chiều và dự đoán đầu ra dựa trên đa số phiếu hoặc trung bình giá trị (Cover & Hart, 1967). Mô hình này thích hợp cho các bài toán

nhận dạng mẫu, như nhận diện khuôn mặt và phân loại văn bản, nhưng hiệu suất giảm với dữ liệu lớn.

2.3.3. Support Vector Machines (SVM)

Support Vector Machines (SVM) là một thuật toán mạnh mẽ dùng cho phân loại và hồi quy. SVM tìm kiếm một siêu phẳng tối ưu để phân tách dữ liệu trong không gian nhiều chiều (Cortes & Vapnik, 1995). Đối với dữ liệu không tuyến tính, SVM sử dụng các hàm kernel như RBF hoặc polynomial để chuyển dữ liệu sang không gian cao hơn, nơi dữ liệu có thể được phân tách tuyến tính. SVM được ứng dụng rộng rãi trong phân loại hình ảnh và phát hiện gian lận tài chính.

2.3.4. Naive Bayes

Naive Bayes là một thuật toán phân loại dựa trên định lý Bayes với giả định rằng các đặc trưng dữ liệu độc lập với nhau. Thuật toán này thường được sử dụng trong các ứng dụng như phát hiện thư rác và phân loại văn bản. McCallum & Nigam (1998) đã chỉ ra rằng mặc dù giả định độc lập không luôn đúng, Naive Bayes vẫn hoạt động tốt với dữ liệu thực tế.

2.3.5. Decision Tree

Decision Tree là một mô hình phân loại và hồi quy được xây dựng dựa trên cấu trúc cây, nơi mỗi nút đại diện cho một điều kiện và các nhánh biểu diễn kết quả của điều kiện đó (Quinlan, 1986). Mô hình này chia dữ liệu thành các nhóm nhỏ hơn dựa trên quy tắc tối ưu hóa thông tin. Decision Tree dễ hiểu và trực quan, nhưng dễ bị overfitting nếu cây quá phức tạp.

2.3.6. Random Forest

Random Forest là một mô hình tập hợp (ensemble) dựa trên nhiều cây quyết định, được tạo ra từ các mẫu dữ liệu ngẫu nhiên (Breiman, 2001). Kết quả của mô hình được xác định bằng cách lấy trung bình (đối với hồi quy) hoặc đa số phiếu (đối với phân loại). Random Forest hoạt động hiệu quả với dữ liệu lớn và giảm thiểu overfitting so với Decision Tree.

2.3.7. Perceptron

Perceptron là một thuật toán học máy tuyến tính đơn giản, được thiết kế để phân loại dữ liệu tuyến tính (Rosenblatt, 1958). Mô hình điều chỉnh trọng số thông qua các lần lặp để tối ưu hóa khả năng phân loại. Perceptron là nền tảng của mạng nơ-ron nhân tạo hiện đại, nhưng hiệu suất của nó bị hạn chế đối với dữ liệu không tuyến tính.

2.3.8. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) là một mô hình học sâu, mô phỏng hoạt động của não bộ con người. ANN sử dụng nhiều lớp ẩn để học các đặc trưng phức tạp

trong dữ liệu (LeCun et al., 2015). Mạng nơ-ron nhân tạo thường được sử dụng trong nhận diện hình ảnh, xử lý ngôn ngữ tự nhiên, và dự đoán tài chính.

2.3.9. Relevance Vector Machine (RVM)

Relevance Vector Machine (RVM) là một mô hình học máy dựa trên lý thuyết Bayesian, được phát triển để cải thiện hiệu quả của SVM (Tipping, 2001). RVM không chỉ đạt độ chính xác cao mà còn cho kết quả phân loại với số lượng vector liên quan ít hơn, giúp giảm chi phí tính toán. RVM thường được sử dụng trong nhận diện mẫu và dự đoán chuỗi thời gian.

CHƯƠNG 3: XỬ LÝ DỮ LIỆU

3.1. Acquire Data

Đầu tiên, nhóm tiến hành thay thế tên cột nhằm thuận tiện hơn trong việc xử lý dữ liệu.

```
df.rename(columns={
    'local_ref_1': 'loc1',
    'vn_marital_status': 'vn_mar',
    'resid_province': 'resid_p',
    'resid_district': 'resid_d',
    'resid_wards': 'resid_w',
    'birth_incorp_date': 'birth_d',
    ' amount_week ': 'amt_w',
    'max_amount_week': 'max_amt_w',
    'min_amount_week': 'min_amt_w',
    'distinct_payment_code_week': 'dist_pay_w',
    'count_payment_code_week': 'cnt_pay_w',
    'distinct_trans_group_week': 'dist_trans_w',
    'distinct_ref_no_week': 'dist_ref_w',
    'amount_month': 'amt_m',
    'max_amount_month': 'max_amt_m',
    'min_amount_month': 'min_amt_m',
    'distinct_payment_code_month': 'dist_pay_m',
    'count_payment_code_month': 'cnt_pay_m',
    'distinct_trans_group_month': 'dist_trans_m',
    'distinct_ref_no_month': 'dist_ref_m',
    'amount_3month': 'amt_3m',
    'max_amount_3month': 'max_amt_3m',
    'min_amount_3month': 'min_amt_3m',
    'distinct_payment_code_3month': 'dist_pay_3m',
    'count_payment_code_3month': 'cnt_pay_3m',
    'distinct_trans_group_3month': 'dist_trans_3m',
    'distinct_ref_no_3month': 'dist_ref_3m',
    'most_act_mar2021_count': 'most_act_m21_cnt',
    'most_act_mar2021': 'most_act_m21',
    'total_act_mar2021': 'total_act_m21',
    ' total_amt_mar2021 ': 'total_amt_m21',
    'most_act_juin2021_count': 'most_act_j21_cnt',
    'most_act_juin2021': 'most_act_j21',
    'total_act_juin2021': 'total_act_j21',
    'total_amt_juin2021': 'total_amt_j21',
    'rd_id': 'rd_id',
    'savingValueMar2021_heoSo': 'saving_m21_HS',
```

```
    'savingValueJuin2021_heoSo': 'saving_j21_HS',
    'totalLoginMar2021_heoSo': 'total_login_m21_HS',
    'totalLoginJuin2021_heoSo': 'total_login_j21_HS',
    'totalSavings2021_heoSo': 'total_savings_21_HS',
    'balanceJuin2021': 'bal_j21',
    'nominal_interestJuin2021': 'nom_int_j21',
    'real_interestJuin2021': 'real_int_j21',
    'nhomno_xhtdJuin2021': 'nhomno_j21',
    'categoryJuin2021': 'cat_j21',
    'sub_productJuin2021': 'sub_prod_j21',
    'loaikyhanJuin2021': 'term_j21',
    'sectorJuin2021': 'sec_j21',
    'product_codeJuin2021': 'prod_j21'
}, inplace=True)
```

Tiếp theo, nhóm xóa bỏ các trường dữ liệu không ghi nhận nợ.

Kết quả thu được tập dữ liệu bao gồm 29.956 dòng và 50 cột.

```
cond = ~df['nhomno_j21'].isna()
df = df.loc[cond]
```

Cuối cùng, nhóm thay đổi các nhóm nợ từ các giá trị 1, 2, 3, 4, 5 thành 0 và 1 (trong đó nhóm 1 và 2 sẽ thành 0, còn nhóm 3, 4, 5 sẽ thành 1).

```
df['nhomno_j21'] = df['nhomno_j21'].map({1: 0, 2: 0, 3: 1, 4: 1, 5: 1})
```

3.2. Analyzing by Describing Data

Đầu tiên, nhóm nghiên cứu tiến hành **xác định kiểu dữ liệu của các cột**.

```
df.dtypes
```

Biến Categorical là những biến có các giá trị được phân loại thành các nhóm hoặc lớp rời rạc, thường không mang tính số học, ta xác định được các cột biến Categorical bao gồm:

- Nomial (không thể sắp xếp thứ tự): loc1, vn_mar, most_act_m21, most_act_j21
- Ordinal (có thể sắp xếp thứ tự): term_j21

Biến Numerical là các biến có dạng số và có thể tính toán, ta xác định được các cột biến Numerical, bao gồm:

- Continuous (Liên tục): amt_w, max_amt_w, min_amt_w, amt_m, max_amt_m, min_amt_m, amt_3m, max_amt_3m, min_amt_3m, total_amt_m21, total_amt_j21, saving_m21_HS, saving_j21_HS, bal_j21, nom_int_j21, real_int_j21.
- Discrete (Rời rạc): resid_p, resid_d, resid_w, most_act_m21_cnt, dist_pay_w, dist_trans_w, dist_ref_w, cnt_pay_w, dist_pay_m, dist_trans_m, dist_ref_m, cnt_pay_m, dist_pay_3m, dist_trans_3m, dist_ref_3m,cnt_pay_3m,

total_act_m21, total_act_j21, total_login_m21_HS, total_login_j21_HS, total_savings_21_HS, rd_id,nhomno_j21, cat_j21, sub_prod_j21, sec_j21, prod_j21.

Tiếp theo, nhóm tiến hành **xác định các cột có giá trị null, khoảng trắng hoặc rỗng** trước khi đưa ra phương án xử lý phù hợp.

```
null_percentage = df.isnull().sum() / len(df) * 100
print(null_percentage)
```

Dữ liệu thuộc nhóm sử dụng dịch vụ Heo Số và nhóm giao dịch theo tuần có giá trị NULL > 35%, nên nhóm quyết định bỏ qua các biến này.

```
df = df.drop([ 'max_amt_w', 'min_amt_w', 'dist_pay_w', 'dist_trans_w', 'dist_ref_w', 'cnt_pay_w', 'amt_w',
 'saving_m21_HS', 'saving_j21_HS', 'total_login_m21_HS', 'total_login_j21_HS', 'total_savings_21_HS'], axis = 1)
```

3.3. Wrangle Data

3.3.1. Xử lý các giá trị không hợp lệ

Nhóm tính toán độ tuổi bằng cách lấy sự chênh lệch giữa năm 2016 và năm sinh, sau đó sử dụng cột độ tuổi 'age' thay thế cho cột 'birth_d' đối với biến nhân khẩu học., theo đó loại bỏ những người có độ tuổi vô lý (dưới 18 tuổi và trên 100 tuổi).

```
# Tính tuổi
df['age'] = (2016 - df['birth_d'])
df=df.drop(['birth_d'], axis=1)
#Loại bỏ những người có độ tuổi vô lý (dưới 18 tuổi và trên 100 tuổi)
df = df[(df['age'] >= 18) & (df['age'] <= 100)]
```

3.3.2 Xử lý null

Các giá trị NaN trong cột tình trạng hôn nhân (vn_mar) sẽ được thay thế bằng mode (giá trị xuất hiện nhiều nhất) của các nhóm kết hợp Tuổi (age) và Giới tính (loc1) để gia tăng độ chính xác, giảm nhiễu ngẫu nhiên cho mô hình.

```
# Tính mode cho mỗi nhóm Giới tính và Độ tuổi
mode_marital = df.groupby(['loc1', 'age'])['vn_mar'].agg(lambda x: x.mode()[0])

# Sử dụng map để điền giá trị thiếu
df['vn_mar'] = df.apply(
        lambda row: mode_marital.loc[(row['loc1'], row['age'])] if pd.isna(row['vn_mar']) else row['vn_mar'],
    axis=1
)

)
```

Ta có thể thấy được không có giá trị ghi nhận ở các giao dịch trong 1 tháng, điều này có nghĩa là khách hàng không thực hiện bất kỳ giao dịch nào trong khoảng thời gian đó. Nhóm quyết định thay thế giá trị NULL bằng 0.

```
#biến đổi null của các biến giao dịch trong 1 tháng
df['dist_pay_m'] = df['dist_pay_m'].replace(np.nan, 0)
df['cnt_pay_m'] = df['cnt_pay_m'].replace(np.nan, 0)
df["dist_ref_m"] = df["dist_ref_m"].replace(np.nan, 0)
```

Biến Sản phẩm tín dụng mà khách hàng sử dụng (sub_prod_j21) có 30% giá trị NULL, trong khi đó biến Nhóm sản phẩm tín dụng mà khách hàng sử dụng (cat_j21) không có giá trị NULL nào, từ đó, nhóm đặt giả định điền giá trị NULL theo giá trị mode (giá trị xuất hiện nhiều nhất) dựa theo biến cat_j21. Cụ thể, đối với mỗi nhóm cat_j21, giá trị NULL trong sub_prod_j21 sẽ được thay thế bằng giá trị mode của sub_prod_j21.

```
df.groupby(['cat_j21'])['sub_prod_j21'].agg(lambda x: x.mode().iloc[0] if not x.mode().empty else np.nan)
```

Với các nhóm sản phẩm có đầu 17 đều không có giá trị mode, đồng nghĩa rằng không có bất kỳ giá trị nào từ các bản ghi có cùng nhóm sản phẩm tín dụng. Nhóm giả định nguyên nhân rằng:

(1) Nhóm sản phẩm tín dụng không có dữ liệu sản phẩm tín dụng thực tế.

(2) Lỗi trong quá trình phân loại hoặc nhập liệu.

Vì vậy, nhóm quyết định điền giá trị NULL bằng 0, nghĩa là sản phẩm tín dụng "Không xác định".

```
df['sub_prod_j21'] = df['sub_prod_j21'].replace(np.nan, 0)
```

Nhóm loại hai biến loại giao dịch khách hàng thực hiện nhiều nhất trong tháng 3 năm 2021 ('most_act_m21') và loại giao dịch khách hàng thực hiện nhiều nhất trong tháng 6 năm 2021 ('most_act_j21') khỏi tập dữ liệu đào tạo vì cho rằng hai biến này không góp phần vào điểm tin cậy của khách hàng.

```
df = df.drop(['most_act_m21','most_act_j21'], axis=1)
```

Với các biến còn lại, nhóm sử dụng mô hình k-Nearest Neighbor Imputation (kNN) để lấp đầy các trường dữ liệu trống.

```
#Khai báo các tham số của KNN
imputer = KNNImputer(n_neighbors=5, weights='uniform', metric='nan_euclidean')
#fit dữ liệu vào mô hình
df_fill= imputer.fit_transform(df[['amt_m','amt_3m','max_amt_m', 'max_amt_3m', 'min_amt_m', 'min_amt_3m', 'dist_trans_m']])
```

```
# điền các dữ liệu còn trống
df['amt_m'] = df_fill[:, 0]
df['amt_3m'] = df_fill[:, 1]
df['max_amt_m'] = df_fill[:, 2]
df['max_amt_3m'] = df_fill[:, 3]
df['min_amt_m'] = df_fill[:, 4]
df['min_amt_3m'] = df_fill[:, 5]
df['dist_trans_m'] = df_fill[:, 6]
```

Cuối cùng, các tài khoản không có thông tin trong năm 2021 được coi là không giao dịch.

```
df['most_act_m21_cnt'].replace(np.nan, 0, inplace= True)
df['total_act_m21'].replace(np.nan, 0, inplace= True)
df['total_amt_m21'].replace(np.nan, 0, inplace= True)
df['most_act_j21_cnt'].replace(np.nan, 0, inplace= True)
df['total_act_j21'].replace(np.nan, 0, inplace= True)
df['total_amt_j21'].replace(np.nan, 0, inplace= True
```

3.3.3 Xử lý các giá trị ngoại lệ

Xử lý các giá trị âm không hợp lệ trong total_act_m21, total_act_j21 (Tổng số giao dịch tháng 3, 6) & total_amt_m21, total_amt_j21 (Tổng giá trị giao dịch tháng 3,6:).

```
    # Kiểm tra và xử lý ngoại lệ ở total_amt_m21 (Tổng giá trị giao dịch trong tháng 3)
    # Kiểm tra các giá trị bất thường
    outliers_amt = df[df['total_amt_m21'] < 0]  # Không mong muốn có giá trị âm
    if not outliers_amt.empty:
        print("Outliers in total_amt_m21 found and removed.")
        df = df[df['total_amt_m21'] >= 0]  # Loại bỏ giá trị âm
    # Kiểm tra và xử lý ngoại lệ ở total_act_j21 (Tổng số giao dịch trong tháng 6)
    outliers_act_j21 = df[df['total_act_j21'] < 0]  # Không mong muốn có giá trị âm
    if not outliers_act_j21.empty:
        print("Outliers in total_act_j21 found and removed.")
        df = df[df['total_act_j21'] >= 0]  # Loại bỏ giá trị âm
```

```
    # Kiểm tra và xử lý ngoại lệ ở total_amt_j21 (Tổng giá trị giao
dịch trong tháng 6)
    outliers_amt_j21 = df[df['total_amt_j21'] < 0]  # Không mong muốn
có giá trị âm
    if not outliers_amt_j21.empty:
        print("Outliers in total_amt_j21 found and removed.")
        df = df[df['total_amt_j21'] >= 0]  # Loại bỏ giá trị âm
```

3.3.4 Chuẩn hóa dữ liệu

Dữ liệu dạng số hầu hết đều thuộc khoảng (0;1), hoặc độ biến thiên không lớn, cho thấy bộ dữ liệu gần như đã được chuẩn hóa sẵn.

CHƯƠNG 4: EXPLORATORY DATA ANALYSIS

4.1. Correlating (Tương quan)

Nhóm tập trung phân tích mức độ tương quan giữa các đặc điểm trong dữ liệu và điểm tin cậy của khách hàng. Mục tiêu là xác định những yếu tố có ảnh hưởng đáng kể đến điểm tin cậy. Các phân tích sơ bộ này sẽ được đối chiếu với kết quả tương quan thu được từ các mô hình được xây dựng sau trong dự án.

Để dễ dàng hơn trong quá trình phân tích, nhóm chia tập dữ liệu thành 5 nhóm để phân tích:

```
#Nhóm 1: Nhân khẩu học
df1=df[['loc1','vn_mar','resid_p','resid_d','resid_w',
'age','nhomno_j21']]
#Nhóm 2: Giao dịch hàng tháng
df2=df[['amt_m',      'max_amt_m',      'min_amt_m',      'dist_pay_m',
'dist_trans_m', 'dist_ref_m', 'cnt_pay_m','nhomno_j21']]
#Nhóm 3: Giao dịch theo quý
df3=df[['amt_3m',     'max_amt_3m',     'min_amt_3m',     'dist_pay_3m',
'dist_trans_3m', 'dist_ref_3m', 'cnt_pay_3m','nhomno_j21']]
#Nhóm 4: Hành vi trong năm 2021
df4=df[['total_amt_m21',        'total_amt_j21',        'total_act_m21',
'total_act_j21', 'most_act_j21_cnt', 'most_act_m21_cnt','nhomno_j21']]
#Nhóm 5: Các chỉ số tài chính
df5=df[['bal_j21',                                      'nom_int_j21',
'real_int_j21','cat_j21','sub_prod_j21','term_j21','sec_j21','prod_j21'
,'nhomno_j21']]
```

4.1.1. Nhóm nhân khẩu học

(1) Giả định:

1. **Giới tính (loc1)** có thể phản ánh sự khác biệt trong hành vi tài chính hoặc chính sách vay vốn. Tuy nhiên nhóm cho rằng không có sự liên kết rõ ràng giữa **giới tính (loc1)** và **tỷ lệ mắc nợ xấu (nhomno_j21)**
2. Với biến **Tỉnh thành resid_p**, các khu vực có môi trường kinh tế ổn định hơn có thể dẫn đến hành vi tín dụng đáng tin cậy hơn.
3. Các biến về quận/huyện/xã có thể gây nhiễu cho mô hình do số lượng phân loại quá lớn, và khác biệt ở từng tỉnh thành.
4. Về **Độ tuổi age**, độ tuổi từ 20-35 tuổi thường là giai đoạn khách hàng có tiềm năng trả nợ tốt hơn do đang trong giai đoạn phát triển sự nghiệp. Trong khi ở độ tuổi 40-50 tuổi, có thể do chi phí gia đình hoặc các khoản vay lớn hơn trở thành gánh nặng.
5. Đối với **Tình trạng hôn nhân vn_mar**, nhóm đã **kết hôn** có thể được xem là ổn định hơn về tài chính do có trách nhiệm gia đình, nguồn thu nhập kép hoặc khả năng tiết

kiệm cao hơn. Trong khi nhóm **độc thân** có thể cần được đánh giá kỹ hơn, nhất là khi họ có thu nhập thấp hoặc công việc không ổn định

(2) Tiến hành:

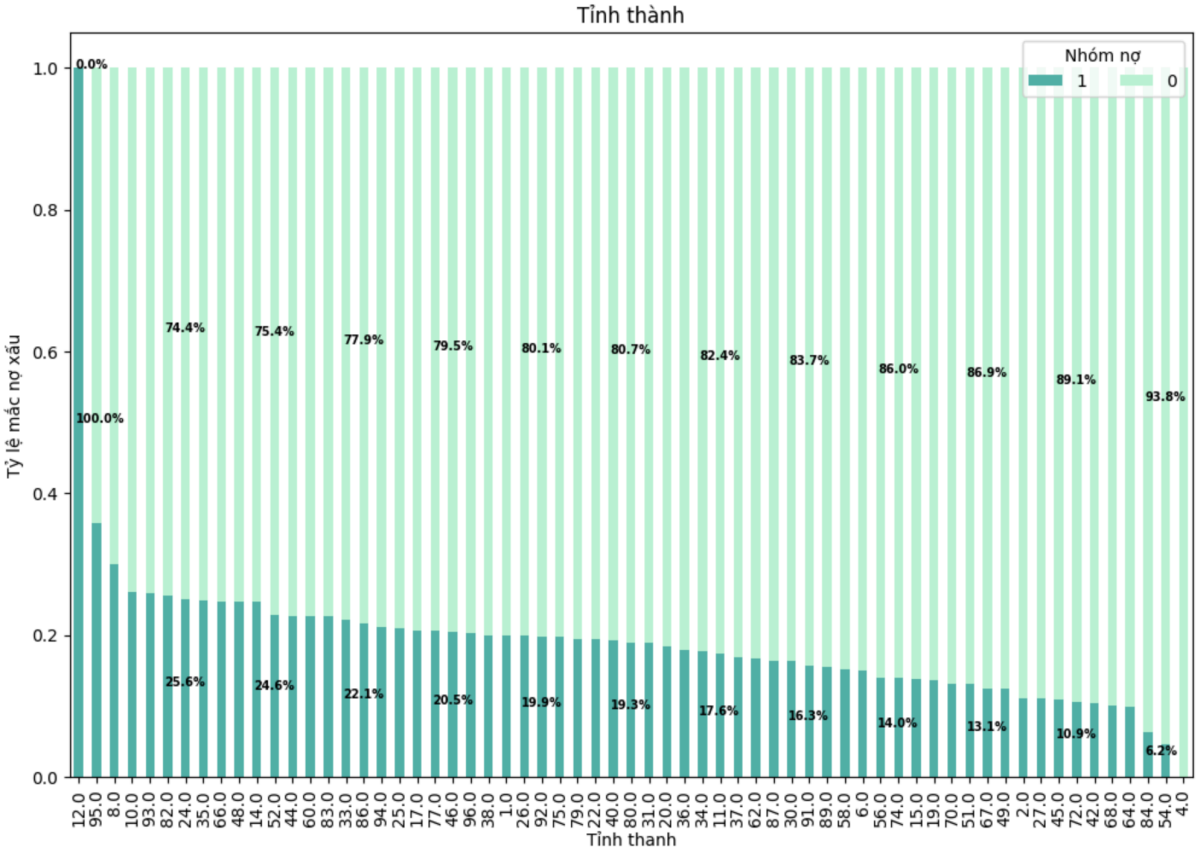
1. Giới tính:

- Tính Pivot Table để kiểm tra xác suất trung bình về độ tin cậy giữa hai giới tính.

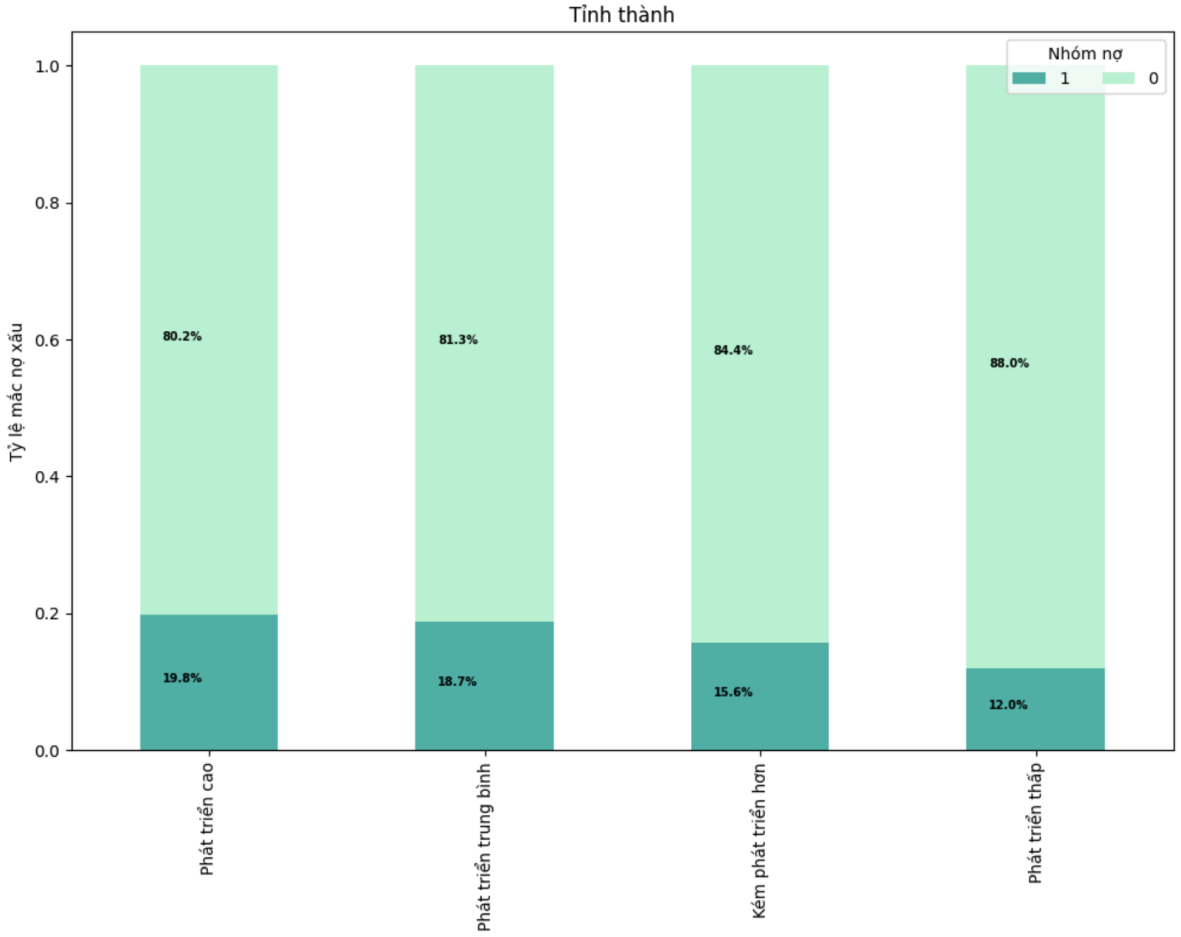
	loc1	nhomno_j21
1	MALE	0.212265
0	FEMALE	0.127271

2. Tỉnh thành sinh sống:

- Tạo biểu đồ Stacked Column Bar để trực quan hóa tỷ lệ mắc nợ xấu theo từng tỉnh thành.

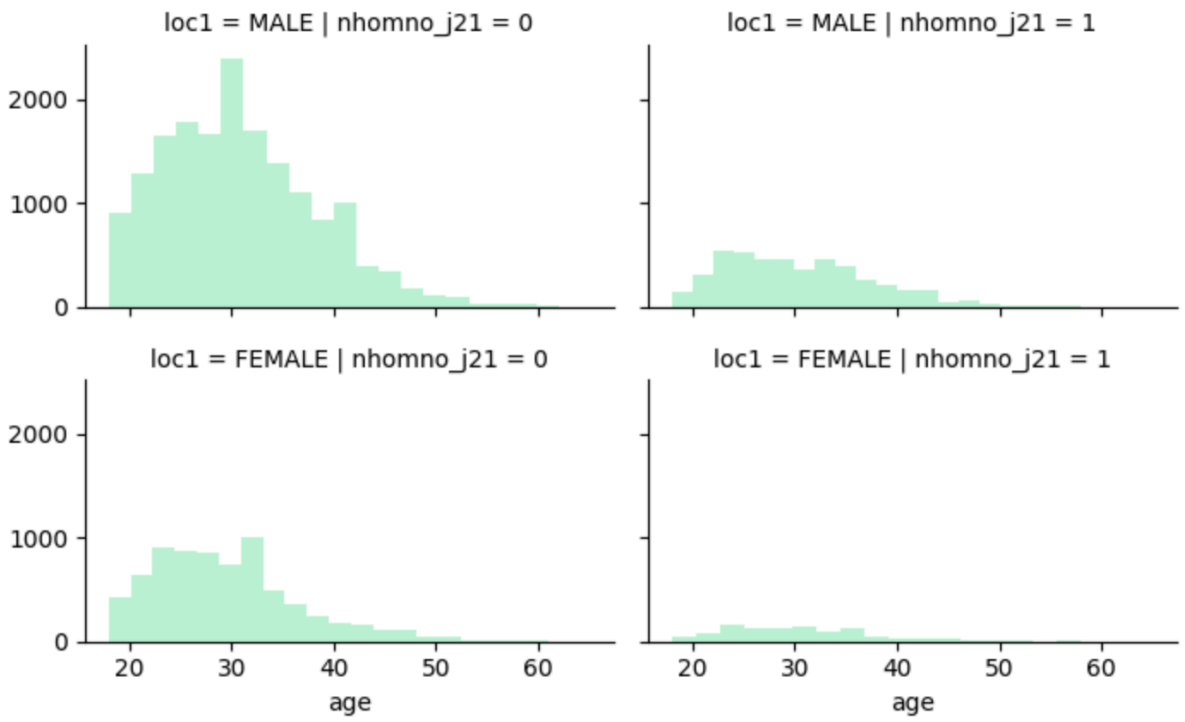


Đối với biến tỉnh thành sinh sống của khách hàng (resid_p), nhóm nhận thấy không có sự chênh lệch rõ rệt giữa các tỉnh thành. Nhóm cho rằng dữ liệu được cấp của biến resid_p là mã số thuế doanh nghiệp. Theo đó, nhóm thay thế tên tỉnh theo bảng của tổng cục thuế cấp. Cuối cùng, nhóm phân loại các tỉnh thành theo mức độ phát triển kinh tế dựa trên thu nhập bình quân của từng tỉnh thành trong năm 2016.



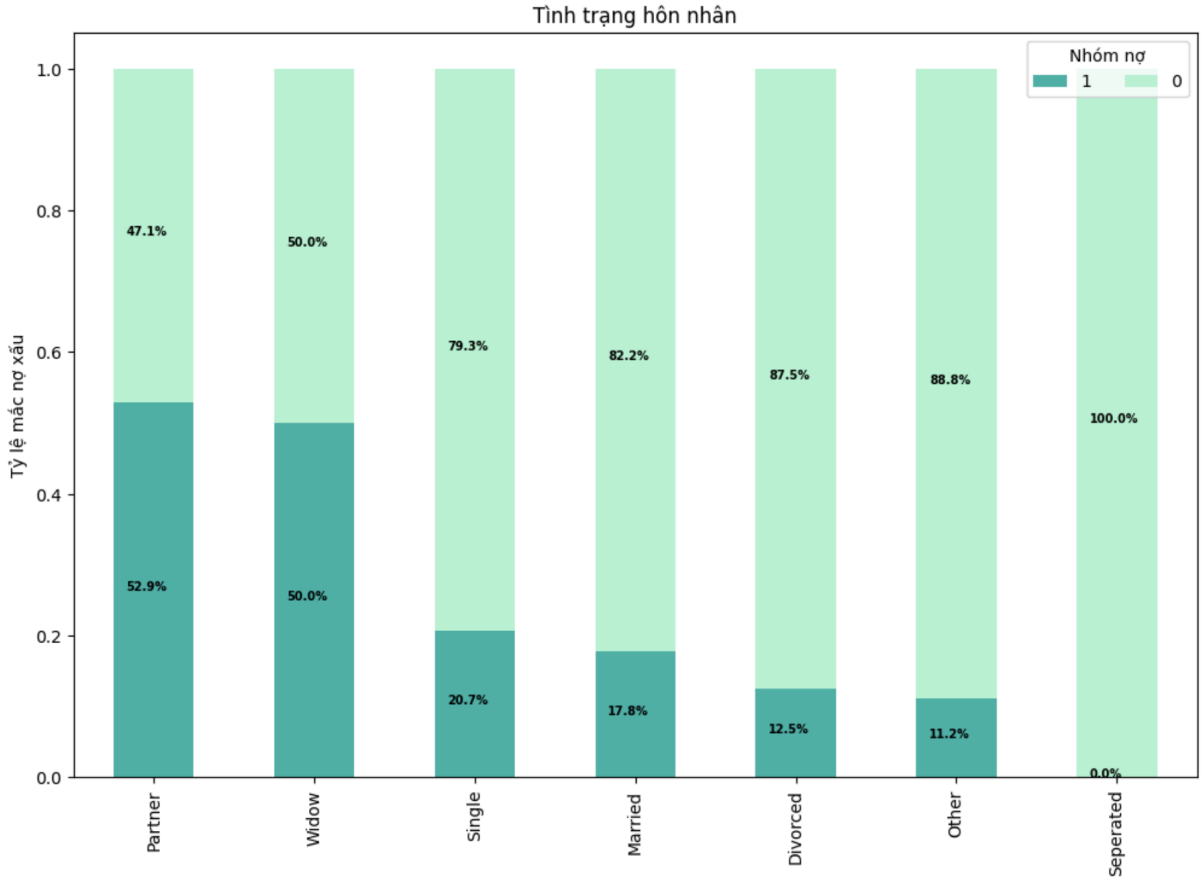
3. Độ tuổi:

- Tạo Histogram kết hợp giữa **Giới tính (loc1) và Độ tuổi (age)** nhằm xem xu hướng hành vi của nhóm khách hàng theo độ tuổi.



4. Tình trạng hôn nhân:

- Sử dụng biểu đồ Stacked Column Bar để xem phân phối về khả năng mắc nợ xấu giữa từng nhóm



(3) Kết quả:

1. Có sự chênh lệch rõ ràng về tỷ lệ nợ xấu giữa các giới tính, tuy nhiên, trên thực tế, không có sự liên kết rõ ràng giữa **giới tính (loc1)** và **tỷ lệ mắc nợ xấu (nhomno_j21)**, thì biến này có thể không cải thiện hoặc thậm chí làm giảm hiệu suất của mô hình.
2. Có sự khác biệt về tỷ lệ nhóm nợ xấu giữa các tỉnh thành ở các mức phát triển (phát triển cao, trung bình, kém, và thấp). Đồng thời, tỷ lệ nhóm nợ xấu giảm dần khi mức độ phát triển tỉnh thành giảm từ cao xuống thấp, cho thấy **Tỉnh thành resid_p** có thể là một xu hướng quan trọng, cho thấy rằng mức độ phát triển của tỉnh thành ảnh hưởng đến rủi ro nợ xấu.
3. Theo biểu đồ Histogram:

- Các khách hàng nam chiếm đa số trong việc vay vốn, điều này cũng có thể giải thích cho lý do khách hàng nam có tỷ lệ mắc nợ xấu cao hơn.
- Xu hướng chung phân phối tuổi của nam và nữ khá tương đồng.
- Những khách hàng mắc nợ xấu có phân phối tuổi thấp, với số lượng tập trung ở độ tuổi dưới 40. Trong khi đó, các khách hàng đáng tin cậy tập trung nhiều ở độ tuổi 20-30. Cho thấy không có sự khác biệt rõ ràng giữa các nhóm biến.

Tuy nhiên, sự khác biệt về quy mô (sample size) vẫn là một yếu tố cần cân nhắc, đặc biệt nếu biến mục tiêu có liên quan đến các nhóm này.

4. Có sự phân hóa rõ ràng giữa biến **tình trạng hôn nhân** trong tỷ lệ mắc nợ xấu của khách hàng. Cho thấy biến vn_mar có thể đóng góp vào việc dự đoán tỷ lệ nợ xấu.

(4) Kết luận:

1. Giữ biến **Tỉnh thành resid_p, độ tuổi age và tình trạng hôn nhân vn_mar** đưa vào mô hình.
2. Drop biến **Giới tính loc1** cùng các biến về hộ khẩu thuộc nhóm quận huyện xã **resid_w**, và **resid_d** ra khỏi mô hình
3. Tạo thêm biến tương tác giữa **Giới tính và Độ tuổi** age* loc1 tăng độ chính xác cho mô hình

4.2.2. Nhóm giao dịch hàng tháng

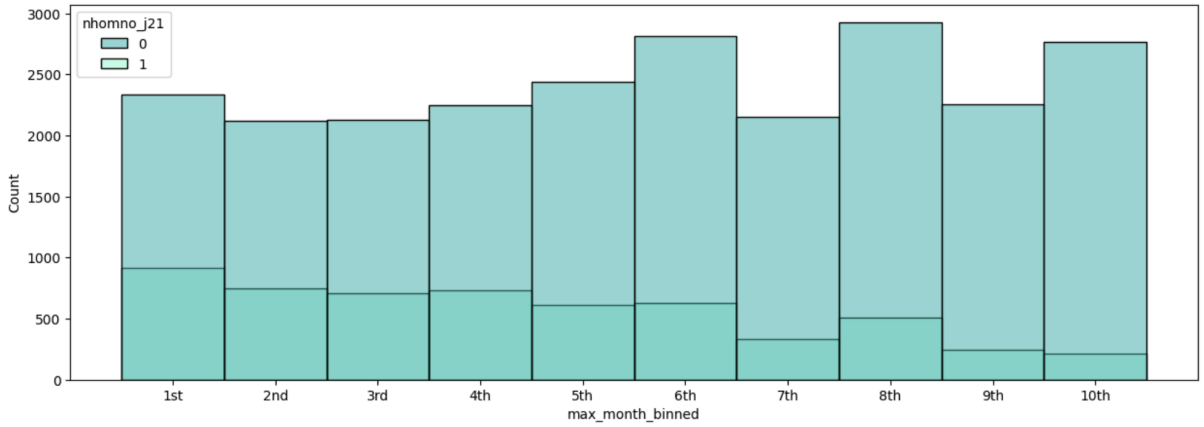
Nhóm xác định mức độ tương quan giữa những biến giao dịch hàng tháng.

Assumptions:

1. Mức độ tương quan giữa các biến giao dịch như min, max, và amount trong các khoảng thời gian (3 tháng, 1 tháng, 1 tuần) là cao và có thể giúp phân tích xu hướng thay đổi qua thời gian.
2. Nhóm khách hàng có mức độ nợ tốt hơn (theo nhomno_j21) thường có tổng số tiền giao dịch tăng dần qua thời gian (3 tháng - 1 tháng) so với các nhóm khách hàng khác.
3. Sự tương quan giữa các biến giao dịch (1 tháng - 3 tháng - 1 tuần) không có tác động đáng kể đến khả năng dự đoán nhóm nợ của khách hàng.
4. Các nhóm khách hàng có mức độ nợ kém hơn có tỷ lệ giao dịch cao nhất và tăng đều đặn qua thời gian (3 tháng - 1 tháng), trái ngược với kết luận rằng nhóm nợ tốt hơn mới có xu hướng này.

Căn cứ từ hình trên, ta rút ra kết luận:

- Mức độ tương quan cao được ghi nhận giữa các cụm giao dịch như min, max và amount trong mọi khoảng thời gian.
- Đặc biệt, các biến giao dịch và dịch vụ trong 1 tháng trước có tương quan mạnh với cả 3 tháng trước và 1 tuần trước, trong khi tương quan giữa 3 tháng và 1 tuần lại thấp hơn. Sự tương quan giữa 1 tháng và 3 tháng là nổi bật nhất. Tiếp theo, nhóm sẽ phân tích sự tăng trưởng tiêu thụ giữa 3 tháng trước và 1 tháng trước để đánh giá mức độ gia tăng trong từng nhóm.



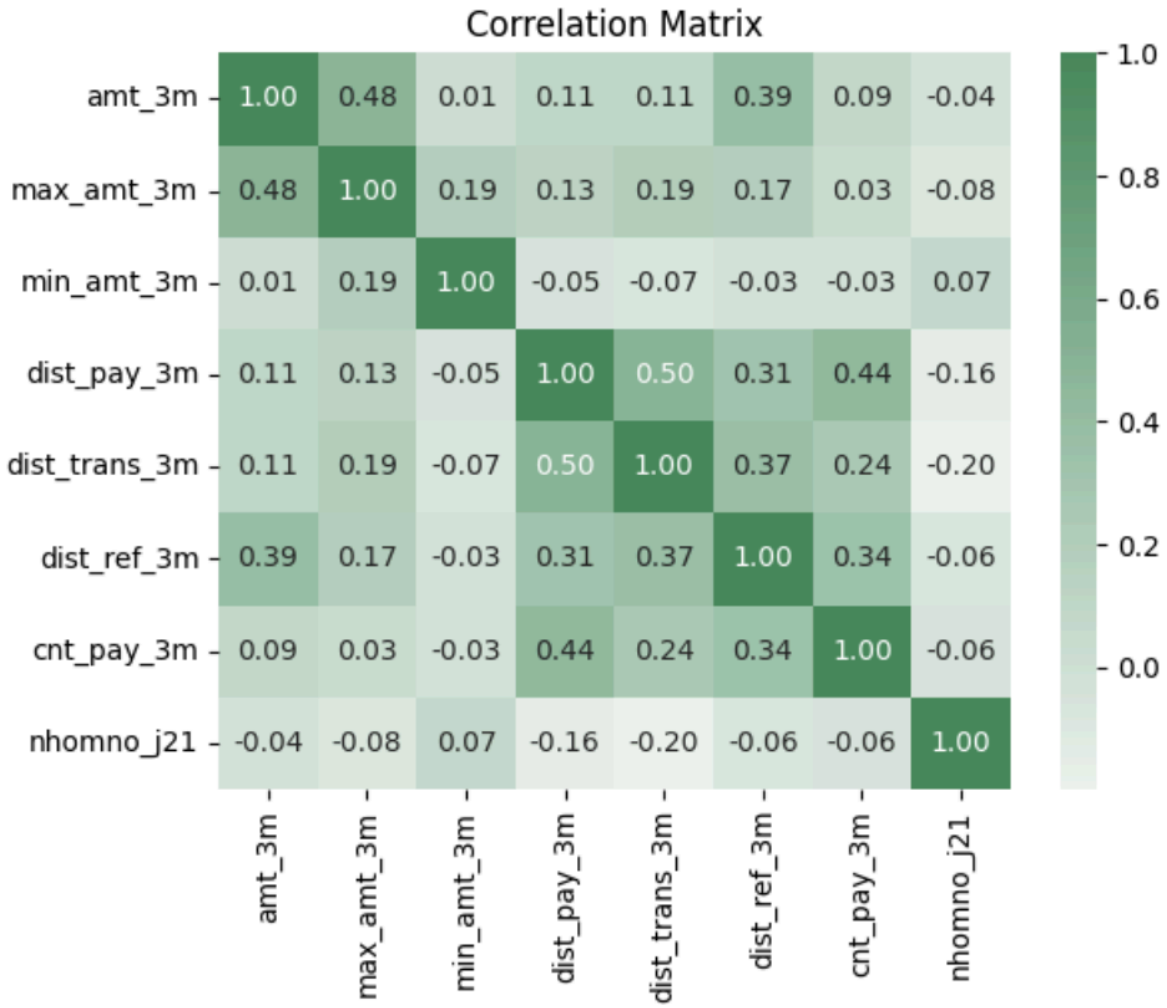
Từ hình trên, ta kết luận:

- Tỷ lệ các khách hàng có tổng số tiền giao dịch tăng lên qua thời gian (3 tháng - 1 tháng) cao hơn ở những nhóm nợ tốt hơn.
- Vì vậy nên đưa vào mô hình biến thay đổi về số tiền giao dịch qua thời gian (3 tháng - 1 tháng) để phân loại và dự đoán biến nhóm nợ.

Kết luận chung: Từ những phân tích phía trên đối với Nhóm 2 (Giao dịch theo tháng), ta kết luận giả định 1 và 2 đúng, giả định 3 và 4 sai

4.2.3. Nhóm giao dịch theo quý

Nhóm xác định mức độ **tương quan giữa những biến giao dịch hàng quý**.



Từ bảng tương quan trên, nhóm xác định được các biến:

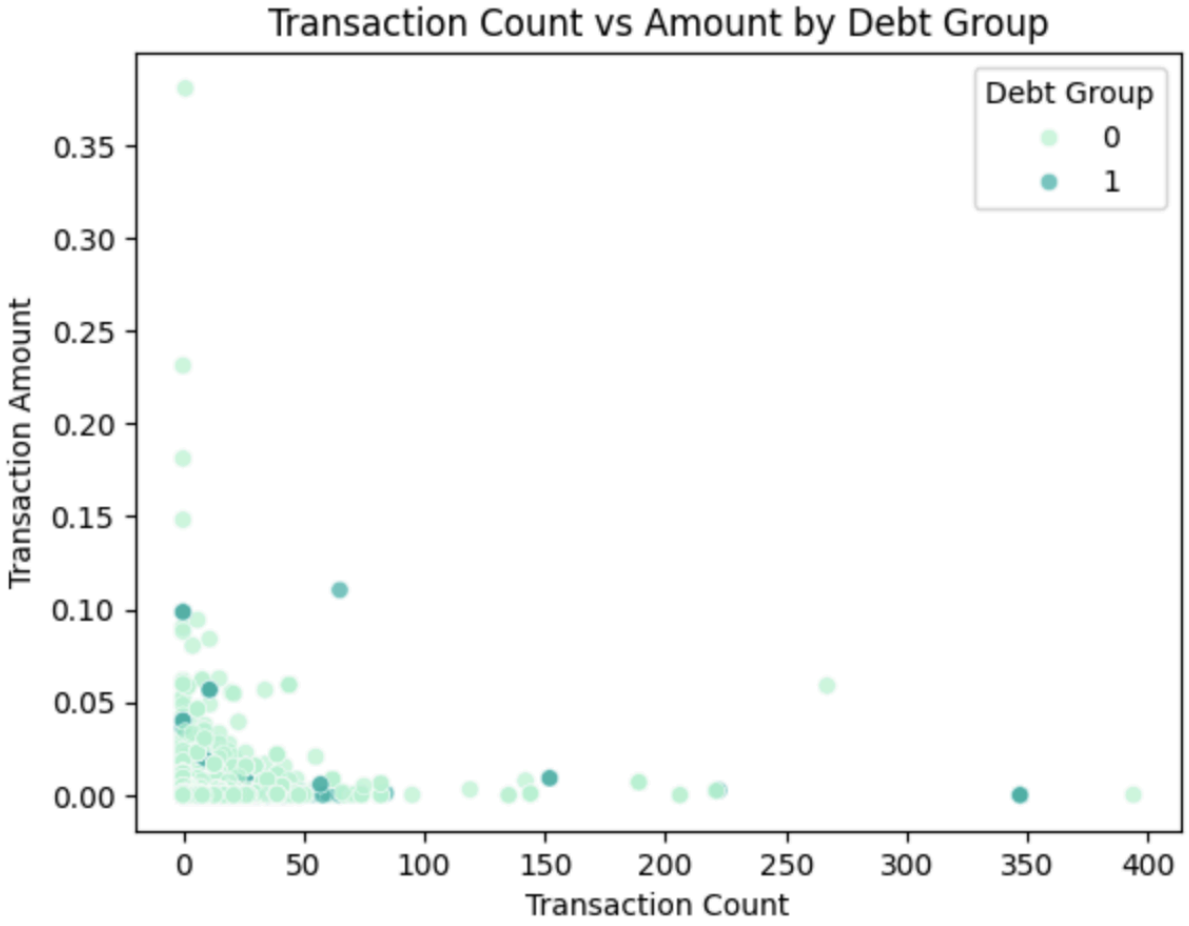
(1) Có tương quan mạnh: amt_3m và max_amt_3m có tương quan vừa phải (0.48), gợi ý rằng tổng số tiền trong 3 tháng gần đây có liên quan đến số tiền lớn nhất trong cùng khoảng thời gian. dist_pay_3m và cnt_pay_3m có tương quan đáng kể (0.44), cho thấy phân phối thanh toán có mối quan hệ với số lần thanh toán.

(2) Có tương quan yếu hoặc không đáng kể: amt_3m và min_amt_3m gần như không có tương quan (0.01). Nhiều biến khác cũng có tương quan rất thấp với min_amt_3m, điều này cho thấy biến này có thể không đóng góp nhiều vào mối quan hệ giữa các biến.

(3) Có mối quan hệ đáng chú ý: dist_ref_3m có tương quan yếu nhưng dương với amt_3m (0.37), max_amt_3m (0.18), và cnt_pay_3m (0.34), gợi ý một mức độ liên hệ nhất định.

Tuy nhiên hầu hết các biến đầu không có tương quan quá mạnh, vì vậy không cần loại biến nào ra.

Tiếp theo, nhóm tìm hiểu về **tương tác giữa nhóm nợ, giá trị giao dịch và số lần giao dịch**.



Ta có thể thấy được:

(1) Nhóm nợ 1 có xu hướng thực hiện giao dịch ít hơn và giá trị nhỏ hơn.

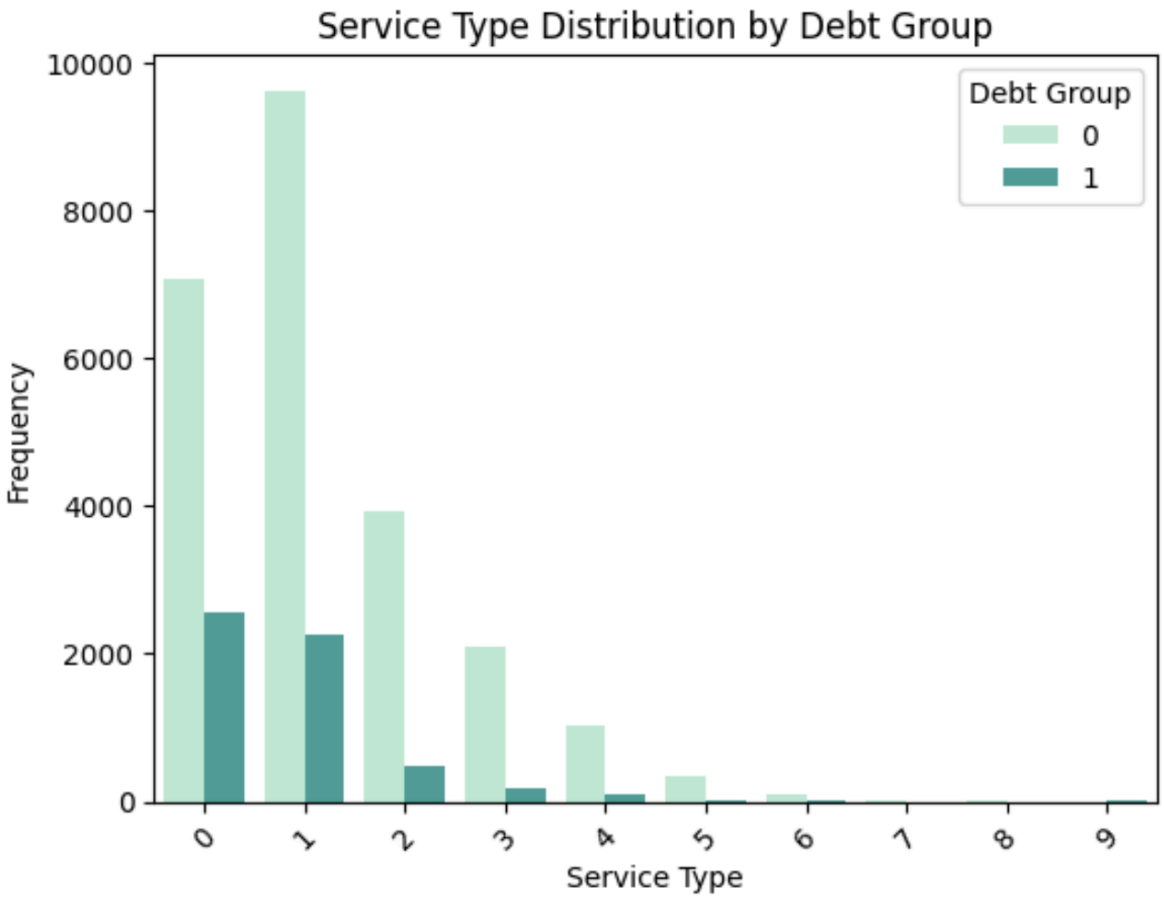
(2) Nhóm nợ 0 chiếm ưu thế cả về số lượng giao dịch lẫn giá trị giao dịch.

Dữ liệu cho thấy sự phân bổ không đồng đều giữa hai nhóm, có thể phản ánh tình trạng tài chính của nhóm nợ 1 yếu hơn.

Tương tác giữa nhóm nợ và loại hình giao dịch

Nhóm nợ 1 có xu hướng tham gia ít giao dịch hơn. Hầu hết các giao dịch tập trung vào loại giao dịch 0, 1 và 2.

Điều này có thể cho thấy nhóm nợ 1 bị hạn chế trong việc tham gia một số loại hình giao dịch, có thể do khó khăn tài chính hoặc các yếu tố rủi ro liên quan.



Tương tác giữa nhóm nợ và nhóm giao dịch

4.2.4. Nhóm hành vi trong năm 2021

(1) Giả định

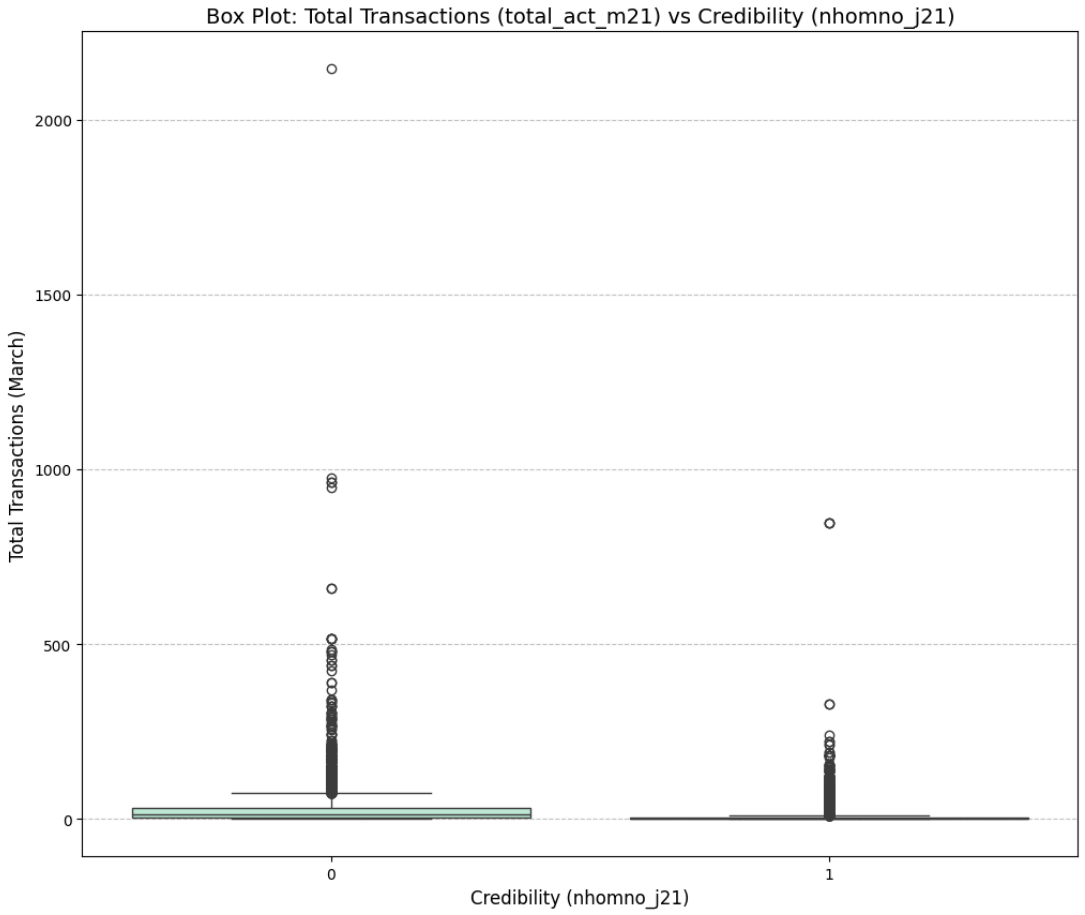
1. **Tổng số giao dịch (total_act_m21) và Tổng giá trị giao dịch (total_amt_m21) cao** thường **đáng tin cậy hơn (credible)** vì họ họ thường xuyên giao dịch hoặc thực hiện các giao dịch giá trị lớn.
2. Biến **Loại Giao Dịch Phổ Biến Nhất (most_act_m21)** không giúp dự đoán biến nợ.
3. Phân tích các biến tháng 3/2021 và kết luận cho cả tháng 6/2021 vì tính đồng nhất của cả 2 nhóm dữ liệu.

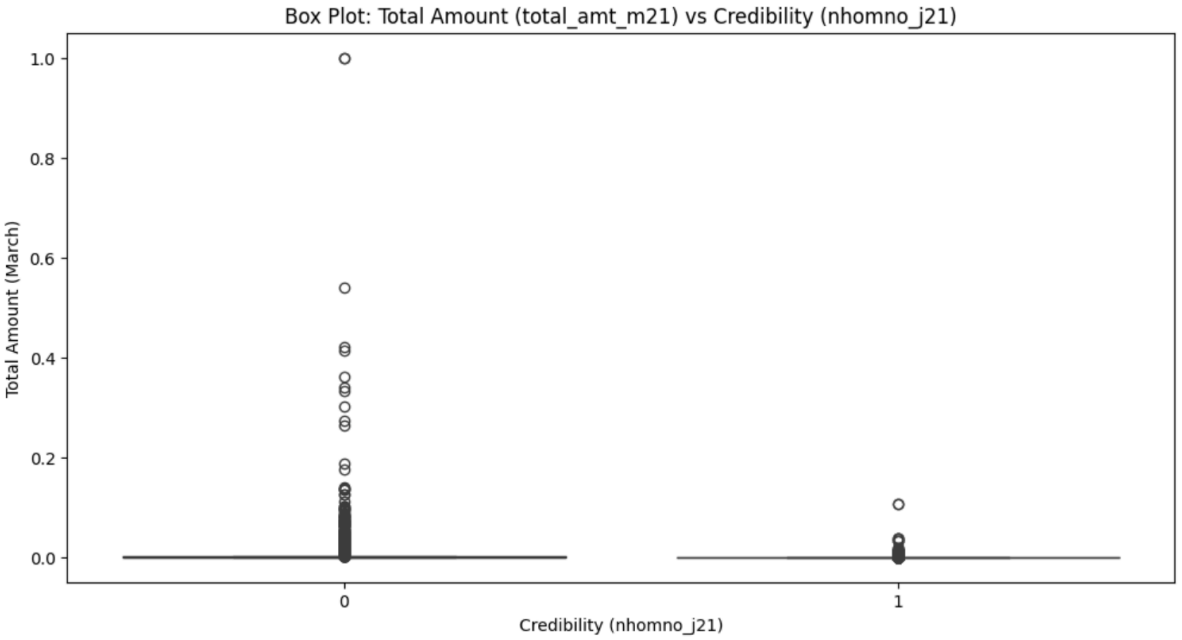
(2) Tiến hành

1. Tổng số giao dịch và tổng giá trị giao dịch:

- Dùng Box Plot để so sánh phân phối **total_act_m21 và total_amt_m21**giữa nhóm **credible (nhomno_j21 = 0) và incredible (nhomno_j21 = 1)**
- Tính Pivot Table để kiểm tra trung bình **total_act_m21** và **total_amt_m21** theo nhóm.

	Credibility (nhomno_j21)	Mean Total Transactions (total_act_m21)	Mean Total Amount (total_amt_m21)
0	0	22.869216	0.001946
1	1	7.702313	0.000246

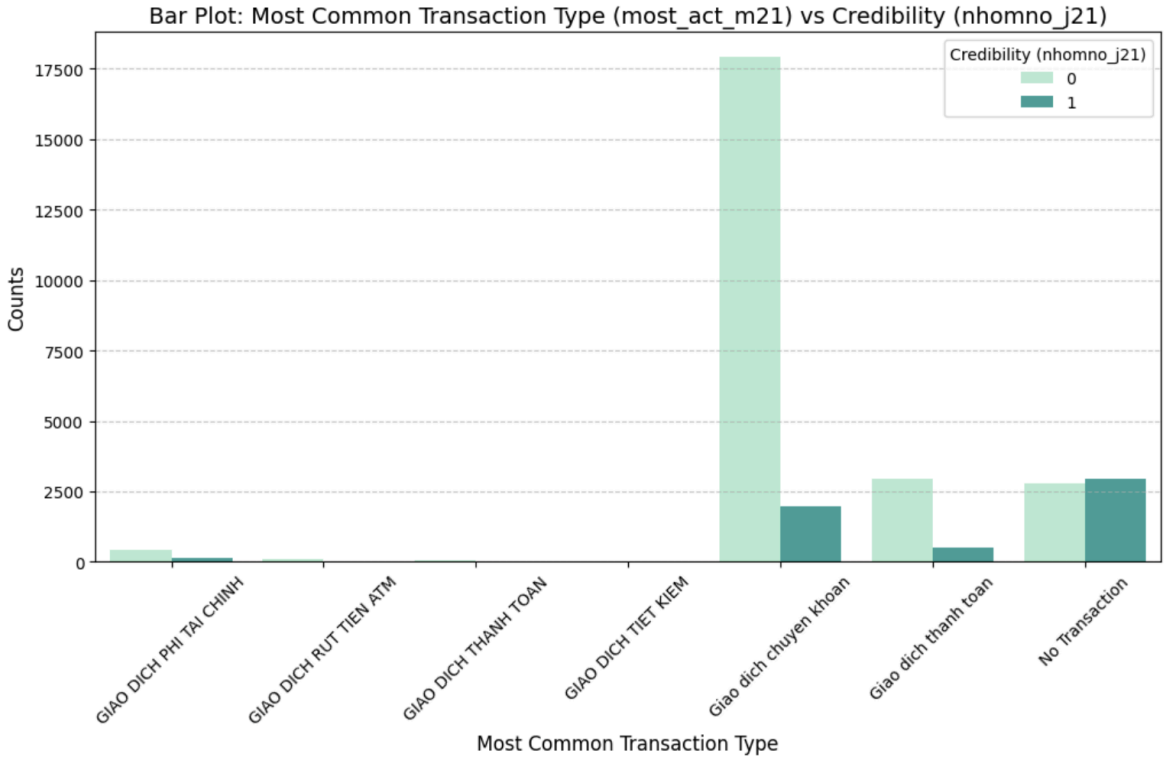




2. Loại giao dịch phổ biến nhất:

- Tạo Pivot Table để kiểm tra loại giao dịch phổ biến nhất **(most_act_m21)** của từng nhóm.
- Sử dụng Bar Plot để trực quan hóa số lượng các loại giao dịch phổ biến nhất theo từng nhóm **nhomno_j21**

	Credibility (nhomno_j21)	Most Common Transaction Type (most_act_m21)
0	0	Giao dich chuyen khoan
1	1	No Transaction



(3) Kết quả

1. Từ Pivot Table, đối với biến Tổng số giao dịch (total_act_m21) và Tổng giá trị giao dịch (total_amt_m21), ta thấy Nhóm credible (nhomno_j21 = 0) có trung bình cao hơn.
2. Từ Box Plot, đối với 2 biến này, ta cũng thấy Total Amount của Nhóm credible (nhomno_j21 = 0) cao hơn.
3. Về biến loại giao dịch phổ biến nhất (most_act_m21) không phân biệt được xu hướng rõ rệt giữa 2 nhóm nợ.

(4) Kết luận

1. Giữ biến Tổng số giao dịch (total_act_m21) và Tổng giá trị giao dịch (total_amt_m21) đưa vào mô hình.
2. Drop biến loại giao dịch phổ biến nhất (most_act_m21, most_act_j21) và biến số lượng giao dịch phổ biến nhất (most_act_m21_cnt, most_act_j21_cnt)

4.2.5. Nhóm các chỉ số tài chính

Phân phối các cột giá trị số

Việc này giúp nhóm xác định, có những hiểu biết ban đầu về dữ liệu của tập dữ liệu. Nhóm biến cat_j21,sub_prod_j21,term_j21,sec_j21,prod_j21, nhomno_j21 bản chất là các biến phân loại được mã hóa, nên nhóm sẽ chỉ phân tích bal_j21, nom_int_j21, real_int_j21.

	bal_j21	nom_int_j21	real_int_j21
count	29805.000000	29805.000000	29805.000000
mean	0.003410	0.005954	0.003283
std	0.016405	0.027083	0.017255
min	0.000000	0.000000	-0.136527
25%	0.000000	0.000000	0.000000
50%	0.000031	0.000099	0.000000
75%	0.001500	0.003571	0.001689
max	1.000000	1.000000	1.000000

Ta có thể thấy được:

- Dữ liệu có thiên hướng bị lệch về phía các giá trị thấp, 75% giá trị đều nhỏ hơn mean (giá trị trung bình).
- Dữ liệu có xu hướng trong khoảng 0-1, cho thấy đây là tập dữ liệu đã được chuẩn hóa.
- Biến tiền gửi có giá trị trung bình khá thấp (so với 1), cho thấy phần lớn các tài khoản có số tiền gửi tương đối thấp so với tổng quy mô.
- Biến lãi suất thực có các giá trị tập trung rất nhỏ với trung vị bằng 0, đồng thời còn có một số giá trị âm, có thể do bối cảnh lạm phát cao hơn lãi suất danh nghĩa.

Phân phối của các cột biến phân loại

	cat_j21	sub_prod_j21	term_j21	sec_j21	prod_j21
count	29805.0	29805.0	29805	29805.0	29805.0
unique	14.0	45.0	3	29.0	35.0
top	21064.0	0.0	NH	1700.0	23231.0
freq	11196.0	9113.0	15965	6871.0	6619.0

Ta có thể thấy được:

- Biến cat_j21 có 14 giá trị duy nhất, cho thấy có 14 nhóm sản phẩm tín dụng. Trong đó, nhóm 21064 là nhóm được sử dụng nhiều nhất với tần suất 37.5%, cho thấy dữ liệu phân bố không đồng đều giữa các nhóm tín dụng.
- Biến sub_prod_j21 có 44 giá trị duy nhất, cho thấy ngoài các sản phẩm tín dụng chưa định nghĩa, có 44 sản phẩm tín dụng được sử dụng, các sản phẩm chưa được định nghĩa lại là là sản phẩm phổ biến nhất. Mặc dù số lượng giá trị duy nhất lớn hơn so với cat_j21, sản phẩm phổ biến nhất vẫn 31% tổng giá trị.
- Trong số 3 loại kỳ hạn vay vốn, Ngắn hạn là hình thức được sử dụng phổ biến nhất, với 51%.
- Trong số 29 ngành kinh doanh, ngành 1700 có tỷ trọng vay vốn nhiều nhất (23%)
- Biến prod_j21gồm 35 giá trị duy nhất, cho thấy có 35 mã sản phẩm được sử dụng và mã 23231 là mã được sử dụng nhiều.

Từ đó, nhóm đặt giả định về quan hệ các biến:

- Khách hàng có số dư tài khoản lớn (bal_j21) có khả năng đáng tin cậy hơn.
- Khách hàng có lãi suất thực tế thấp (real_int_j21) có khả năng trả nợ tốt hơn.
- Khách hàng có loại kỳ hạn khoản vay dài hạn (term_j21) có thể có khả năng trả nợ tốt hơn.

Tương quan của các cột biến phân loại

(1) Giả định:

1. Các biến gồm **Nhóm sản phẩm tín dụng (cat_j21), Sản phẩm tín dụng (sub_prod_j21),Ngành kinh doanh (sec_j21), vàMã sản phẩm vay (prod_j21)`** có thể cho thấy xu hướng và hành vi tài chính của khách hàng trong từng lĩnh vực, qua đó giúp ngân hàng dự đoán khả năng trả nợ đúng hạn.
2. Khách hàng có **Số dư tài khoản (bal_j21) lớn** thường **đáng tin cậy hơn (credible)** vì số dư tài khoản lớn phản ánh khả năng tài chính tốt, nghĩa là khách hàng có thể đáp

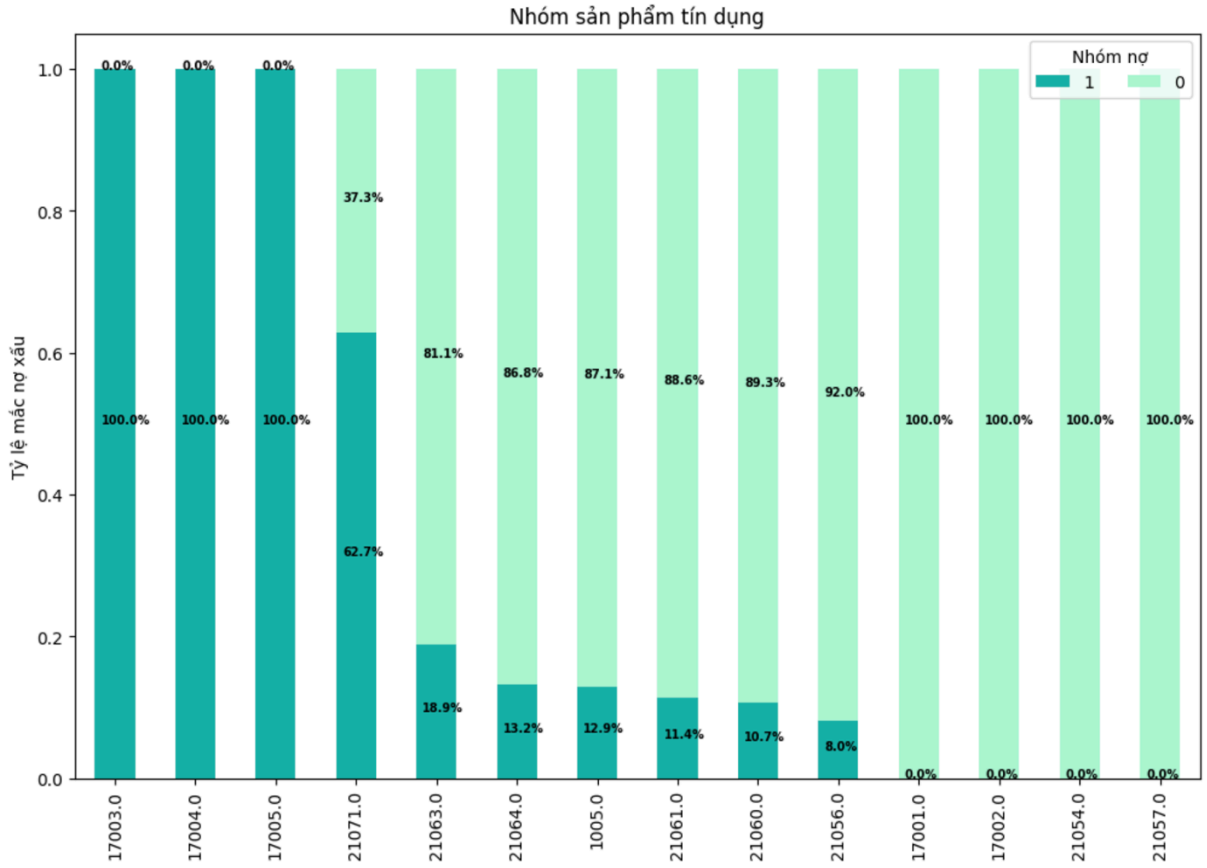
ứng các nghĩa vụ tài chính. Điều này giúp giảm rủi ro cho ngân hàng hoặc đối tác kinh doanh.

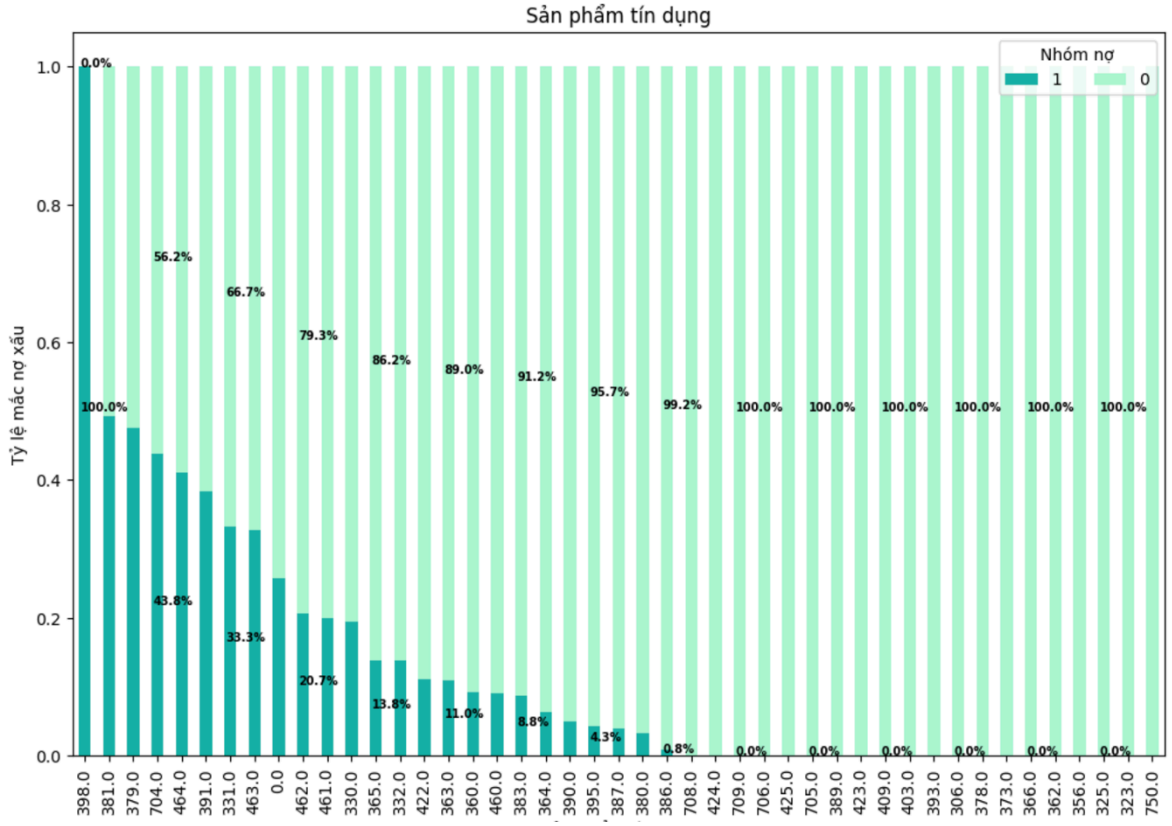
3. **Lãi suất danh nghĩa (nom_int_j21)** thấp giúp khách hàng có khả năng trả nợ đúng hạn do khách hàng vay tiền phải trả ít lãi hơn, làm giảm gánh nặng tài chính.
4. **Kỳ hạn khoản vay (term_j21) dài** giúp khách hàng dễ trả nợ hơn trong môi trường lãi suất thấp, nhưng ngân hàng cần xem xét rủi ro lâu dài, khi khách hàng mất cảnh giác với nghĩa vụ tài chính. Ngược lại **Kỳ hạn khoản vay (term_j21) ngắn** là dấu hiệu tốt cho thấy khách hàng muốn hoàn thành nghĩa vụ tài chính nhanh chóng, thể hiện trách nhiệm cao. Cần kết hợp term_j21 với biến khác để dễ dàng hơn trong phân tích.
5. **Lãi suất thực tế (real_int_j21)** tuy có thể là biến quan trọng trong việc xem xét mối quan hệ, tuy nhiên biến số này còn phụ thuộc vào lạm phát thực tế, có thể gây nhiễu trong việc dự đoán biến mục tiêu

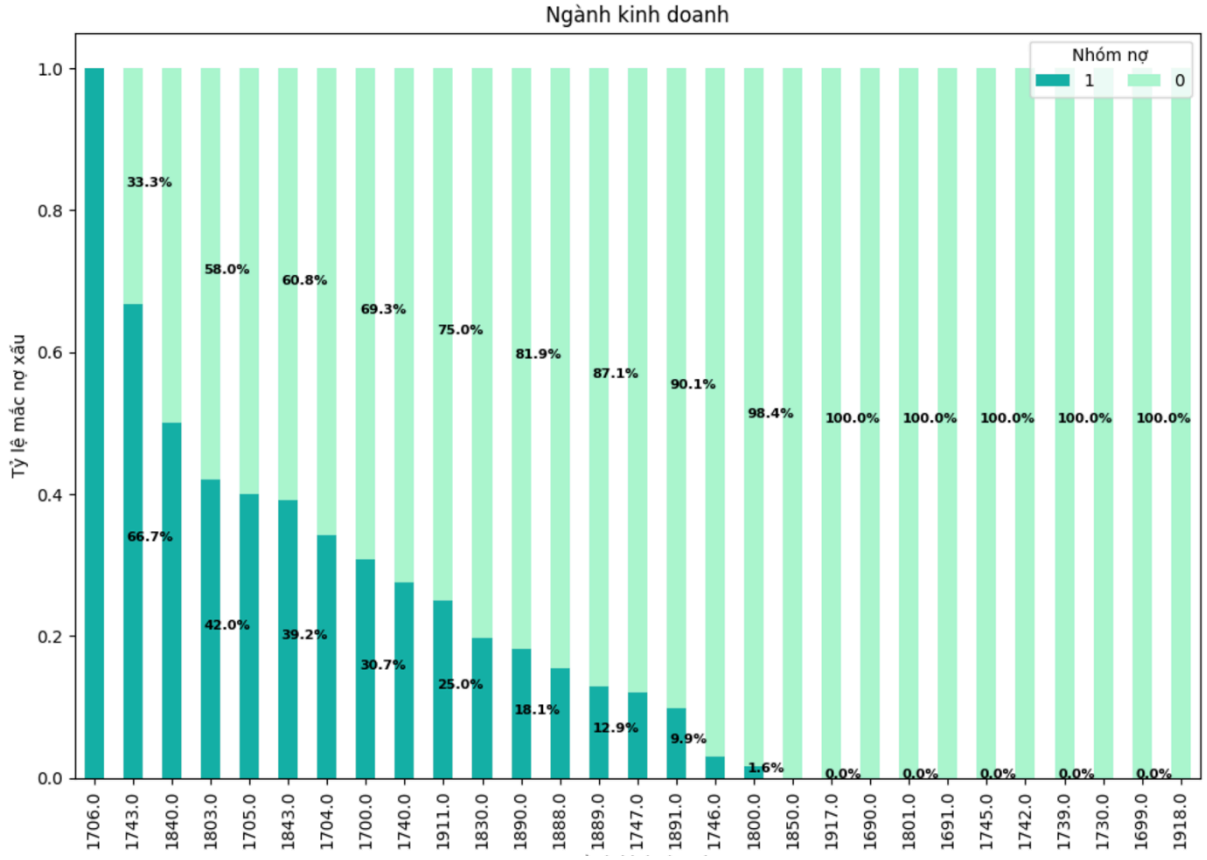
(2) Tiến hành:

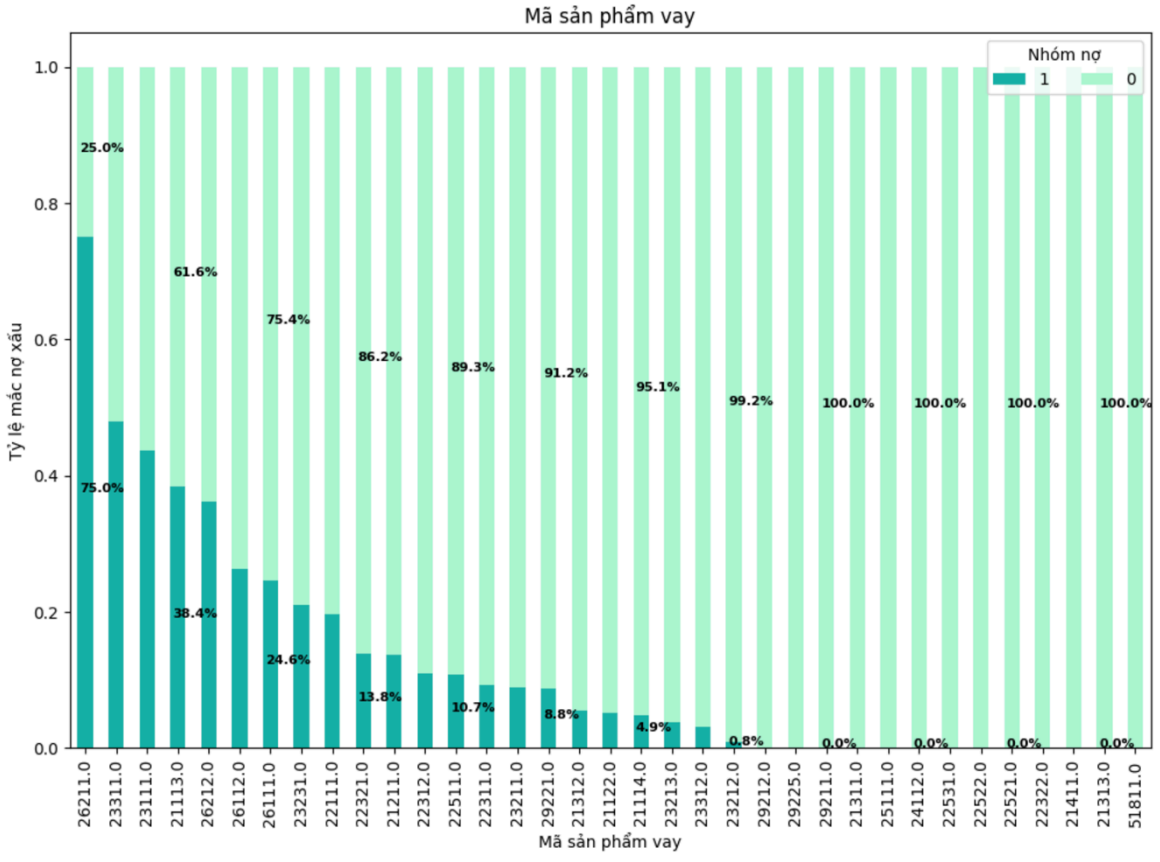
1. **Nhóm sản phẩm tín dụng (cat_j21), Sản phẩm tín dụng (sub_prod_j21), `Ngành kinh doanh (sec_j21), Mã sản phẩm vay (prod_j21), và Kỳ hạn khoản vay (term_j21):**

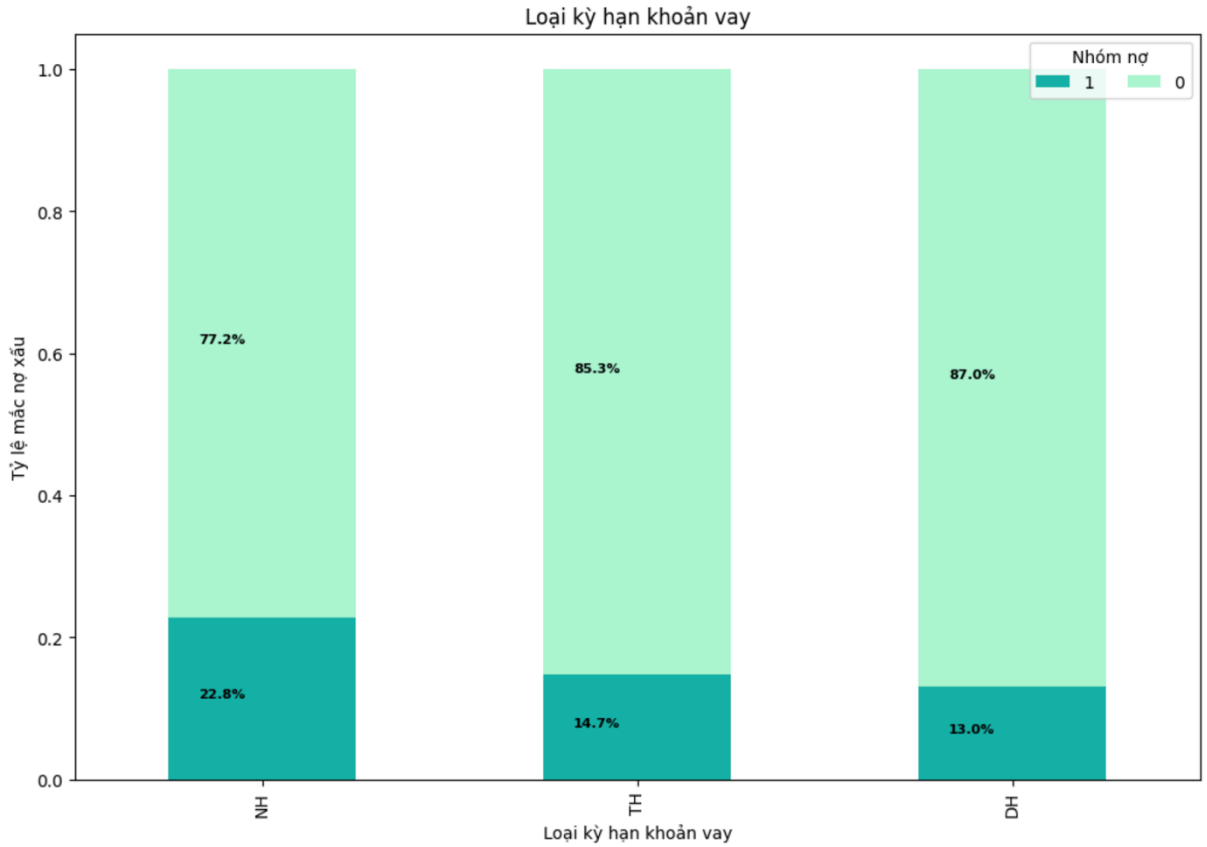
- Nhóm phân tích sự tương quan giữa các đặc tính bằng cách so sánh chúng với nhau thông qua Stacked Column Bar để so sánh giữa nhóm **credible (nhomno_j21 = 0) và incredible (nhomno_j21 = 1)**


Nhóm sản phẩm tín dụng
Nhóm nợ
1
0
Tỷ lệ mắc nợ xấu
0.0
0.2
0.4
0.6
0.8
1.0
0.0%
0.0%
0.0%
37.3%
81.1%
86.8%
87.1%
88.6%
89.3%
92.0%
100.0%
100.0%
100.0%
100.0%
100.0%
100.0%
100.0%
100.0%
62.7%
18.9%
13.2%
12.9%
11.4%
10.7%
8.0%
0.0%
0.0%
0.0%
0.0%
17003.0
17004.0
17005.0
21071.0
21063.0
21064.0
1005.0
21061.0
21060.0
21056.0
17001.0
17002.0
21054.0
21057.0


Sản phẩm tín dụng
Nhóm nợ
1
0
Tỷ lệ mắc nợ xấu
0.0
0.2
0.4
0.6
0.8
1.0
0.0%
56.2%
66.7%
79.3%
86.2%
89.0%
91.2%
95.7%
99.2%
100.0%
100.0%
100.0%
100.0%
100.0%
100.0%
100.0%
43.8%
33.3%
20.7%
13.8%
11.0%
8.8%
4.3%
0.8%
0.0%
0.0%
0.0%
0.0%
0.0%
0.0%
398.0
381.0
379.0
704.0
464.0
391.0
331.0
463.0
0.0
462.0
461.0
330.0
365.0
332.0
422.0
363.0
360.0
460.0
383.0
364.0
390.0
395.0
387.0
380.0
386.0
708.0
424.0
709.0
706.0
425.0
705.0
389.0
423.0
409.0
403.0
393.0
306.0
378.0
373.0
366.0
362.0
356.0
325.0
323.0
750.0


Ngành kinh doanh
Nhóm nợ
1
0
Tỷ lệ mắc nợ xấu
1.0
0.8
0.6
0.4
0.2
0.0
33.3%
66.7%
58.0%
42.0%
60.8%
39.2%
69.3%
30.7%
75.0%
25.0%
81.9%
18.1%
87.1%
12.9%
90.1%
9.9%
98.4%
1.6%
100.0%
0.0%
1706.0
1743.0
1840.0
1803.0
1705.0
1843.0
1704.0
1700.0
1740.0
1911.0
1830.0
1890.0
1888.0
1889.0
1747.0
1891.0
1746.0
1800.0
1850.0
1917.0
1690.0
1801.0
1691.0
1745.0
1742.0
1739.0
1730.0
1699.0
1918.0


Mã sản phẩm vay
Nhóm nợ
1
0
Tỷ lệ mắc nợ xấu
Mã sản phẩm vay
1.0
0.8
0.6
0.4
0.2
0.0
25.0%
75.0%
61.6%
38.4%
75.4%
24.6%
86.2%
13.8%
89.3%
10.7%
91.2%
8.8%
95.1%
4.9%
99.2%
0.8%
100.0%
0.0%
26211.0
23311.0
23111.0
21113.0
26212.0
26112.0
26111.0
23231.0
22111.0
22321.0
21211.0
22312.0
22511.0
22311.0
23211.0
29221.0
21312.0
21122.0
21114.0
23213.0
23312.0
23212.0
29212.0
29225.0
29211.0
21311.0
25111.0
24112.0
22531.0
22522.0
22521.0
22322.0
21411.0
21313.0
51811.0

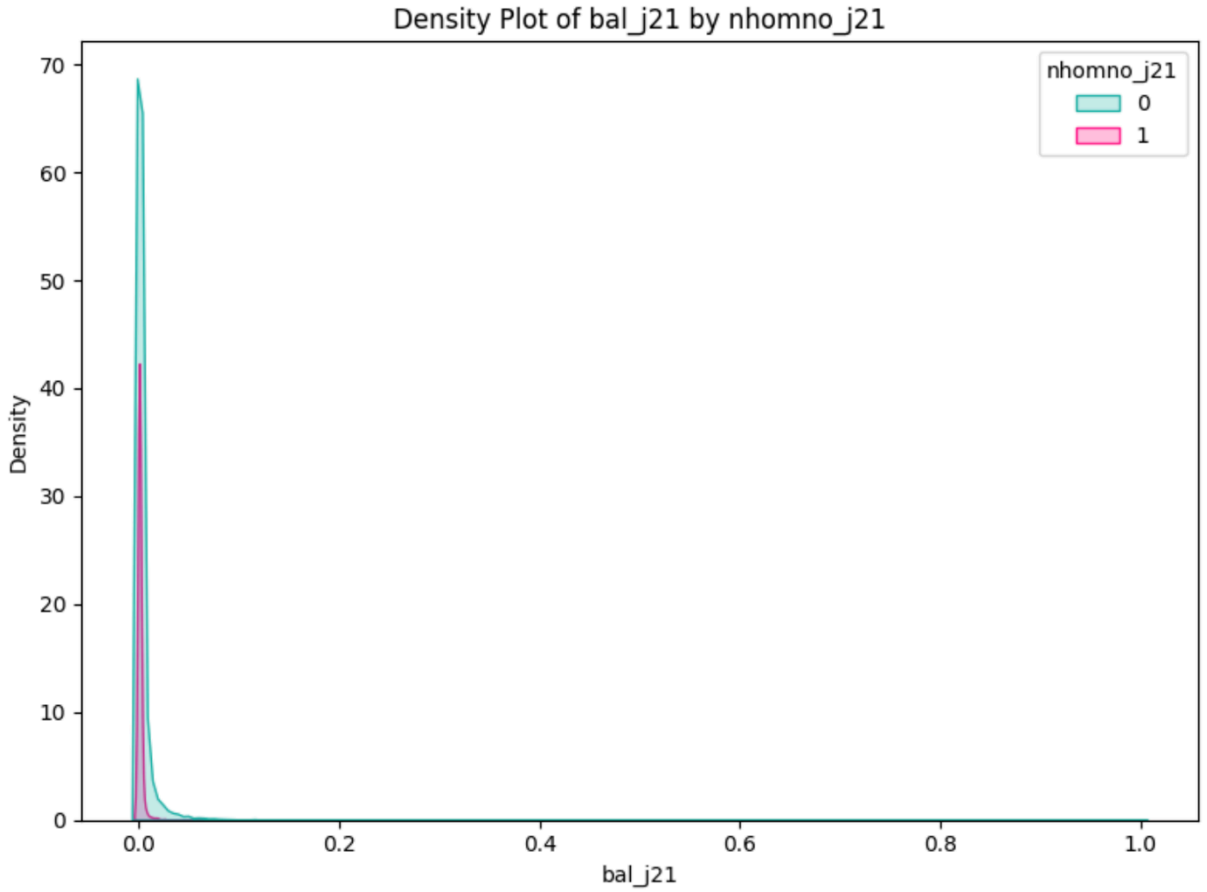


2. **Số dư tài khoản tiền gửi ngân hàng:**

- Biểu đồ histogram hữu ích trong việc phân tích các biến bal_j21 liên tục. Biểu đồ histogram có thể chỉ ra sự phân phối của các mẫu bằng cách sử dụng các nhóm (bins) được xác định tự động hoặc các dải có phạm vi đều nhau.

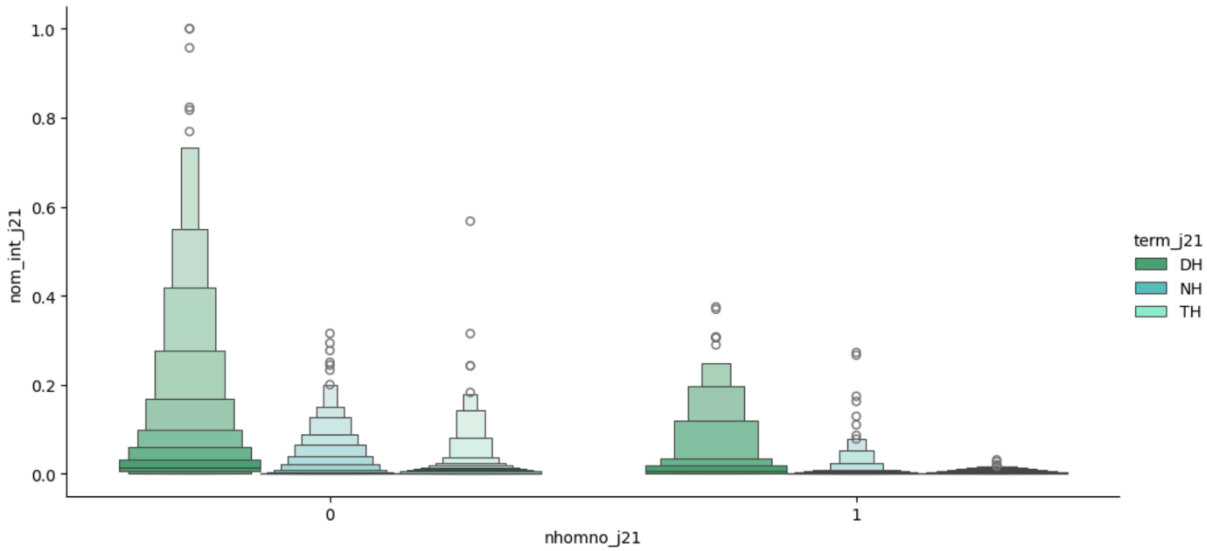
Tuy nhiên, do dữ liệu sử dụng tập trung trong khoảng 0-1, nhóm sử dụng **Density Plot** có thể giúp làm rõ hơn phân phối dữ liệu

Lưu ý rằng trục x trong biểu đồ histogram đại diện cho số lượng mẫu hoặc khách hàng.



3. Lãi suất danh nghĩa và kỳ hạn khoản vay:

- Nhóm sử dụng Boxenplot (viết tắt từ box plot extended) để phân tích mối quan hệ giữa lãi suất danh nghĩa, kỳ hạn khoản vay, và độ tin cậy của khách hàng, nhằm so sánh phân phối và sự khác biệt giữa các nhóm khách hàng.
- Nhóm sử dụng Boxenplot thay vì Boxplot do Boxenplot được thiết kế để xử lý tốt hơn với dữ liệu lớn hoặc phân phối dài, giúp hiển thị rõ các phần tử ở phân vị cao/thấp.



(3) Kết quả:

1. Từ Stacked Bar Column, nhóm nhận thấy tỷ lệ nợ xấu giảm dần theo các **nhóm sản phẩm tín dụng cat_j21, sản phẩm tín dụng sub_prod_j21, ngành kinh doanhsec_j21, và mã sản phẩm vayprod_j21`** với sự phân hóa rõ rệt, thậm chí xuất hiện mối tương quan tuyệt đối (0.00, 1.00)

- Tuy nhiên, khi nhìn vào dữ liệu, nhóm nhận thấy rằng sự biến thiên giữa các sản phẩm có thể không phải là ngẫu nhiên mà có sự tương tác giữa các nhóm sản phẩm. Do sự chênh lệch theo các nhóm nhất định. Vì vậy, nhóm quyết định bổ sung một biến tương tác giữa nhóm sản phẩm tín dụng và sản phẩm tín dụng cụ thể vào mô hình để xem xét liệu sự kết hợp giữa các yếu tố này có ảnh hưởng đến tỷ lệ nợ xấu hay không. Việc này có thể giúp mô hình trở nên chính xác hơn và cải thiện khả năng dự đoán.
- Tương tự với Ngành kinh doanh và Mã sản phẩm vay

2. Không có sự phân hóa rõ ràng giữa **các loại kỳ hạn term_j21** trong phân tích tỷ lệ mắc nợ xấu, cho thấy kỳ hạn không có sự đóng góp rõ ràng trong việc dự đoán.
3. Nhóm nhận thấy mức nợ xấu chỉ xuất hiện ở các tài khoản có số dư tài khoản ở mức thấp. Điều này chứng tỏ biến ***Số dư tài khoản tiền gửi*** có thể giúp đánh giá xác suất mức tín dụng tốt hơn. Nhóm quyến định đưa **Số dữ tài khoản bal_j21** vào mô hình.
4. Theo Boxenplot, nhóm nhận thấy:

- Nhóm 0 (khách hàng đáng tin cậy) có lãi suất vay cao hơn. Điều này cho thấy ngân hàng có thể đang ưu tiên áp dụng lãi suất ưu đãi hơn cho nhóm khách hàng không đáng tin cậy (nhóm 1), để giảm rủi ro mất vốn. Ngược lại, với nhóm đáng tin cậy (nhóm 0), ngân hàng có thể tự tin áp dụng mức lãi suất cao hơn mà không lo ngại khả năng mất vốn, tối đa hóa lợi nhuận. Vì vậy, nhóm quyết định đưa biến **lãi suất danh nghĩa (nom_int_j21)** vào mô hình
- Kỳ hạn của khoản vay dài (DH) có lãi suất cao nhất trong cả hai nhóm nợ, đồng thời xuất hiện nhiều giá trị ngoại lệ hơn, đặc biệt ở nhóm không có nợ xấu. Cho thấy nhóm đáng tin cậy (nhóm 0) có thể bao gồm nhiều khách hàng thuộc phân khúc cao cấp hơn, dẫn đến khoản vay lớn và kèm theo mức lãi suất cao hơn, đặc biệt với kỳ hạn dài, để ngân hàng có thể tối ưu lợi nhuận. Vì vậy, nhóm quyến định đưa **biến tương tác giữa biến kỳ hạn khoản vay và lãi suất danh nghĩa** vào mô hình.

(4) Kết luận:

1. Giữ biến sản phẩm tín dụng cat_j21, sản phẩm tín dụng sub_prod_j21, ngành kinh doanhsec_j21, mã sản phẩm vayprod_j21, số dư tài khoản tiền gửibal_21, vànom_int_j21` đưa vào mô hình.
2. Drop biến kỳ hạn khoản vay term_j21 và real_int_j21
3. Thêm các biến tương tác giữa nhóm sản phẩm tín dụng và sản phẩm tín dụng cat_j21* sub_prod_j21; Ngành kinh doanh và Mã sản phẩm vay sec_j21 *prod_j21

4.2. Wrangle Data

4.2.1. Corrrecting (Loại bỏ hoặc điều chỉnh)

Vì lãi suất thực tế bị ảnh hưởng bởi lạm phát, là yếu tố khách quan bên ngoài và khó có số liệu trong bộ dữ liệu dự đoán tương lai, nhóm quyết định loại bỏ cột real_int_j21.

```
df.drop(['real_int_j21'], axis =1, inplace = True)
```

Loại bỏ các cột biến loc1, resid_p,resid_w, resid_d khỏi mô hình.

```
df.drop(['loc1', 'resid_w', 'resid_d'], axis=1, inplace=True)
```

Nhóm đưa biến tình trạng hôn nhân vn_mar về dạng số theo phương pháp One Hot Encoding.

```
df = pd.get_dummies(data=df,
                    columns=["vn_mar"],
                    prefix="OHE",
                    prefix_sep="_",
                    drop_first=True)
```

Tiếp theo, cột biến Loại kỳ hạn khoản vay term_j21 và Tỉnh thành resid_p (đã phân theo mức độ phát triển) là các biến ordinal. Nhóm mã hóa các biến này theo dải số từ 1, theo thứ tự có tỷ lệ mắc nợ xấu tăng dần.

```
df['term_j21'] = df['term_j21'].map( {'NH': 1, 'TH': 2, 'DH':3}
).astype(int)
df['resid_p']=df1['resid_p'].map({'Phát triển cao':1, 'Phát triển trung
bình':2, 'Kém phát triển hơn': 3, 'Phát triển thấp':4}).astype(int).
```

4.2.2. Creating (Tạo biến mới)

Nhóm tạo biến tương tác giữa cat_j21 và sub_prod_21 như đã phân tích ở trên, trước đó, nhóm chuẩn hóa (Scaling) các giá trị trước khi nhân để tránh gặp phải giá trị lớn do sự khác biệt lớn giữa các giá trị của chúng.

```
scaler = StandardScaler()

df[['cat_j21', 'sub_prod_j21']] = scaler.fit_transform(df[['cat_j21',
'sub_prod_j21']])

df['cat_j21*sub_prod_j21'] = df.cat_j21 * df.sub_prod_j21

df.loc[:, ['cat_j21*sub_prod_j21', 'cat_j21', 'sub_prod_j21']].head()
```

Tạo biến tương tác tương tự giữa sec_j21 và prod_j21 giống 2 biến trên.

```
df['bal_j21*term_j21'] = df.bal_j21 * df.term_j21

df.loc[:, ['bal_j21*term_j21', 'bal_j21', 'term_j21']].head()
```

CHƯƠNG 5: MODEL, PREDICT AND SOLVE

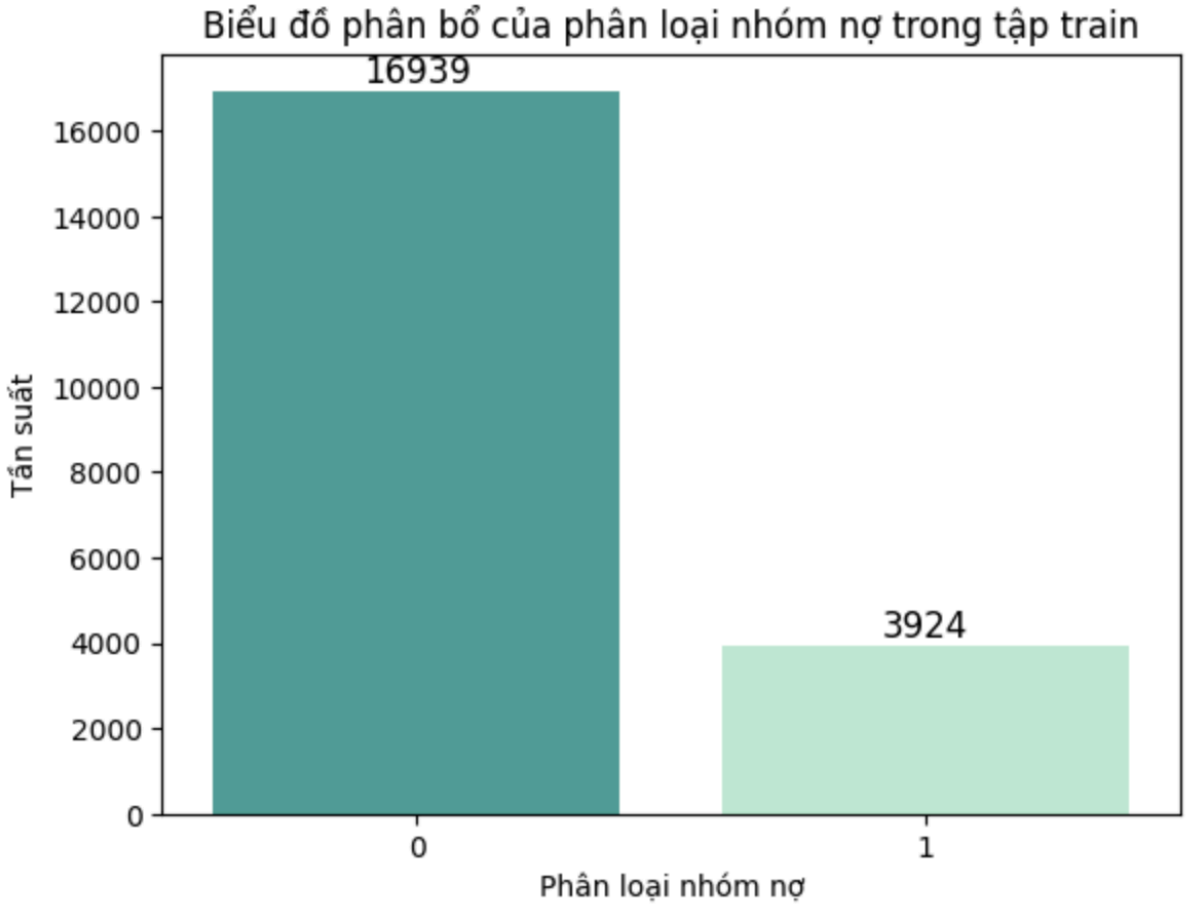
Nhóm chọn cột để stratify thay vì random trong quá trình chia dữ liệu nhằm giúp phân phối đều các nhóm đặc trưng quan trọng (feature-based stratify).

```
X = df.drop(['nhomno_j21'], axis=1)
y = df['nhomno_j21']
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, stratify=df['cat_j21'])
train_df = pd.concat([X_train, y_train], axis=1)
```

5.1. Cân bằng dữ liệu

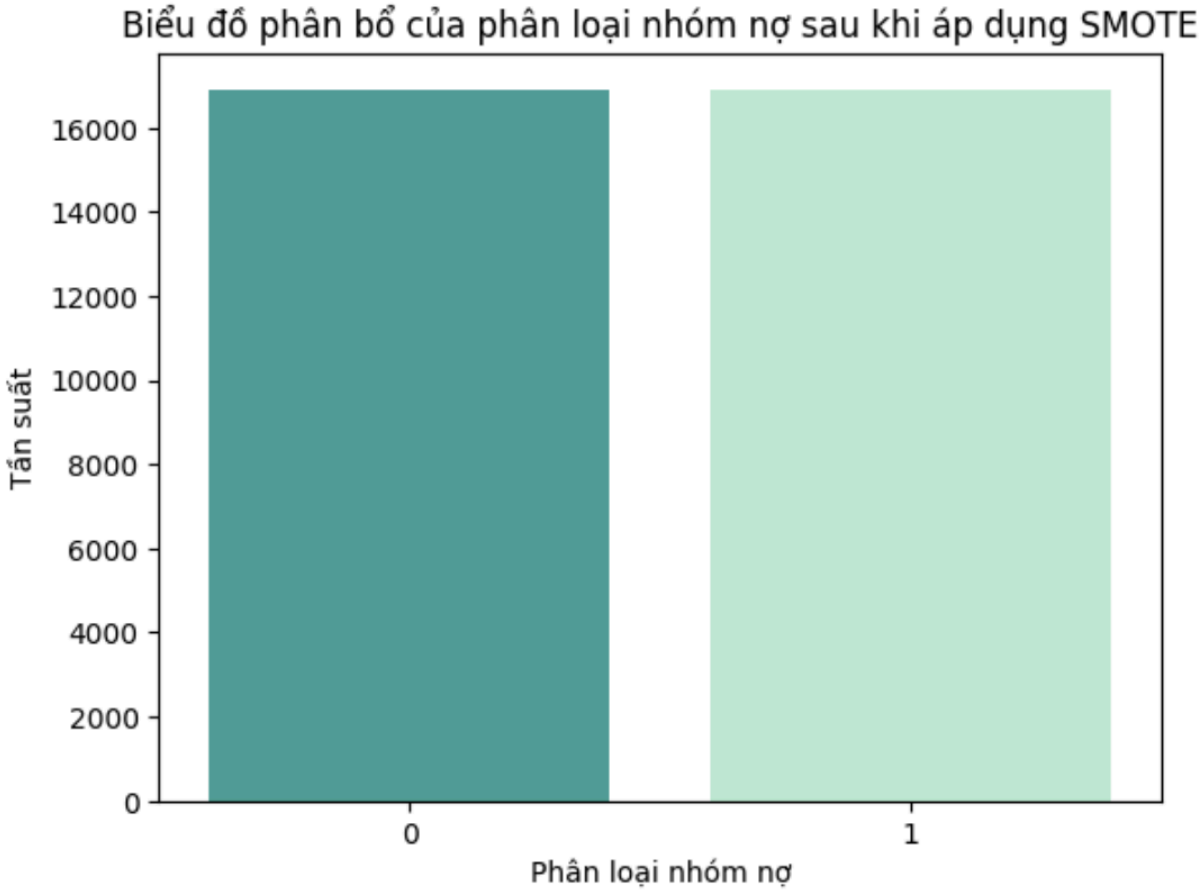
Mất cân bằng dữ liệu (Class Imbalance) là một vấn đề phổ biến trong các bài toán học máy, đặc biệt là trong các bài toán phân loại (classification problems).

Các lớp nhóm nợ trong bộ dữ liệu phân loại có số lượng mẫu không đều, trong đó lớp thuộc nhóm nợ đáng tin cậy chiếm ưu thế hơn đáng kể so với lớp nợ xấu, chiếm 80% số quan sát, điều này có thể khiến ngộ nhận chất lượng mô hình hoặc mô hình sẽ ưu tiên học về lớp chiếm đa số, làm mất nhận diện lớp thiểu số.



Nhóm sử dụng phương pháp Borderline-SMOTE: thay vì lấy k toàn bộ của dữ liệu xung quanh thì lấy k ở tiếp 1 cạnh, các cạnh này nằm ở đường phân chia giữa lớp còn những điểm bị lẫn như ở ví dụ trên sẽ bỏ qua.

```
import math
# Tăng mẫu
def bordersmote(x,y):
    k_neighbors = math.ceil(sum(y) * 0.01)
    m_neighbors = math.ceil(sum(y) * 0.01)

    bordersmote = BorderlineSMOTE(sampling_strategy=1,
                                  k_neighbors=k_neighbors,
                                  m_neighbors=m_neighbors)

    x, y = bordersmote.fit_resample(x, y)

    return x, y

X_train, y_train = bordersmote(X_train, y_train)

from collections import Counter
```



5.2. Model

Mục tiêu của bài toán là dự đoán đầu ra (khách hàng có khả năng mắc nợ xấu hay không) từ các tính năng đầu vào (giới tính, độ tuổi, giao dịch...). Đồng thời, bài

toán này được giải quyết thông qua học có giám sát (supervised learning), trong đó mô hình học từ tập dữ liệu đã có nhãn (được biết kết quả). Với hai tiêu chí này - Học có giám sát cộng với **Phân loại và Hồi quy**, chúng ta có thể thu hẹp lựa chọn mô hình của mình xuống còn một vài mô hình. Các mô hình mà nhóm sẽ sử dụng bao gồm:

- Logistic Regression
- KNN or k-Nearest Neighbors
- Support Vector Machines
- Naive Bayes classifier
- Decision Tree
- Random Forrest
- Perceptron
- Artificial neural network

Sau khi xây dựng các mô hình, nhóm tổng hợp được các chỉ số đánh giá mô hình, cụ thể như sau:

	Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
5	Random Forest	0.961865	0.912615	0.881688	0.896885	0.985799
6	LightGBM	0.944531	0.888089	0.806778	0.845483	0.978692
4	Decision Tree	0.938045	0.808872	0.878121	0.842075	N/A
0	Logistic Regression	0.188101	0.188101	1.000000	0.316642	0.501041
7	Perceptron	0.188101	0.188101	1.000000	0.316642	N/A
1	K-Nearest Neighbors	0.526057	0.198585	0.500595	0.284363	N/A
2	Support Vector Machines	0.468799	0.186555	0.542806	0.277676	0.500436
3	Gaussian Naive Bayes	0.601879	0.184688	0.326992	0.236052	0.50137
8	Artificial Neural Network	0.811899	0.000000	0.000000	0.000000	0.499862

Kết luận: Mô hình tốt nhất: Random Forest.

Đạt được hiệu suất cao nhất về mọi mặt: F1-Score, AUC-ROC, Precision, Recall.

Phù hợp với dữ liệu hiện tại và không yêu cầu nhiều tuning.

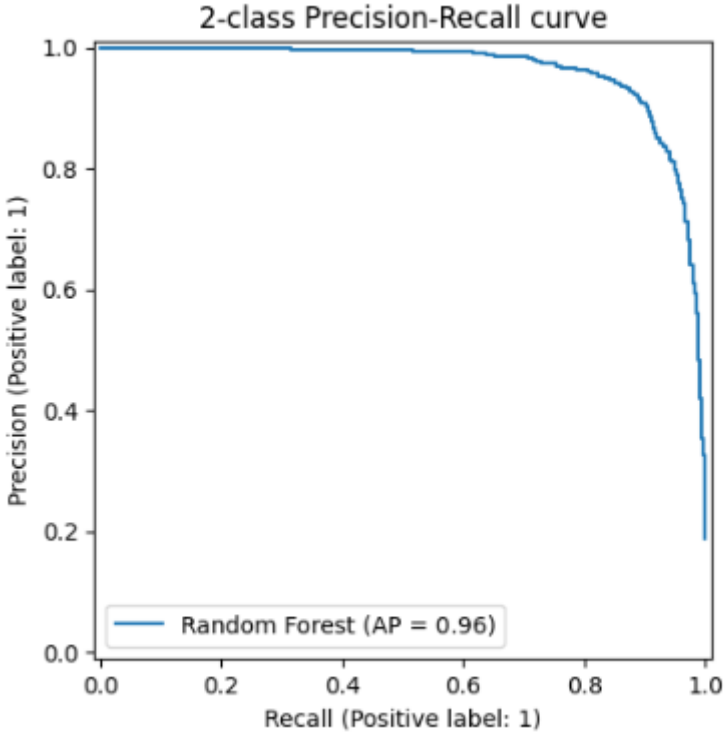
LightGBM:

Là lựa chọn thứ hai, hiệu suất gần tương đương nhưng nhanh hơn Random Forest, phù hợp nếu cần triển khai trên tập dữ liệu lớn.

5.3. Tối ưu hóa Threshold

Mục tiêu của bài toán nhằm dự đoán nợ xấu, từ đó đánh giá độ tin cậy của khách hàng. Chính vì vậy, nhóm ưu tiên phát hiện tất cả các trường hợp thật dương (True Positive), không muốn bỏ sót khách hàng có rủi ro cao. Tuy nhiên Recall của các mô hình chỉ đạt mức cao nhất là 87.52%, nghĩa là cứ 100 khách hàng không đáng tin cậy, nhóm dự đoán sai ít nhất 13 khách hàng, điều này có thể gây ra nhiều rủi ro đối với hệ thống ngân hàng. Nhóm quyết định thay đổi Threshold (Ngưỡng phân loại để cải thiện mô hình).

Nhóm tiến hành vẽ biểu đồ Precision-Recall curve cho thấy sự đánh đổi (trade-off) giữa precision và recall cho các ngưỡng khác nhau.



Chỉ số **Average Precision (AP)** bằng 0.96, gần bằng 1, cho thấy mô hình Random Forest khá tốt trong việc dự đoán.

Sau khi tiến hành điều chỉnh threshold và phân tích hiệu suất, nhóm nhận thấy ngưỡng tốt nhất là Ngưỡng = 0.34, với F1-score = 0.88. Ngưỡng này có sự cân bằng tốt giữa precision (0.81) và recall (0.95). Đồng thời, mô hình vẫn tối ưu hóa được giá trị recall.

Nhóm thu được bảng kết quả sau khi điều chỉnh Threshold:

	Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
0	Random Forest	0.951912	0.828378	0.940201	0.880754	0.988216
1	LightGBM	0.943972	0.833708	0.878626	0.855578	0.980451
2	Decision Tree	0.937374	0.803985	0.883955	0.842076	0.916884
7	Artificial Neural Network	0.189443	0.188990	1.000000	0.317899	0.500345
3	Logistic Regression	0.188884	0.188884	1.000000	0.317750	0.499969
5	Gaussian Naive Bayes	0.188884	0.188884	1.000000	0.317750	0.504902
6	Perceptron	0.188884	0.188884	1.000000	0.317750	0.500000
4	K-Nearest Neighbors	0.516551	0.186279	0.462996	0.265670	0.500497

Kết luận: Mô hình tốt nhất: Random Forest.

Đạt được hiệu suất cao nhất về mọi mặt: F1-Score, AUC-ROC, Precision, Recall.

CHƯƠNG 6: GIẢI PHÁP CHO CÁC NHÓM NỢ XẤU

Giải pháp cho nhóm nợ xấu cần tập trung vào việc cải thiện quy trình thẩm định tín dụng, phân loại khách hàng và áp dụng công nghệ phân tích dữ liệu hiện đại. Ngân hàng cần sử dụng các thuật toán học máy như Random Forest hoặc Logistic Regression để đánh giá xác suất khách hàng rơi vào nhóm nợ xấu, từ đó hỗ trợ quá trình ra quyết định. Bên cạnh đó, áp dụng nguyên tắc 5Cs (Character, Capacity, Capital, Collateral, Conditions) sẽ giúp đảm bảo sự đánh giá toàn diện và chính xác về khả năng tài chính của khách hàng. Đặc biệt, cần chú trọng đến tài sản thế chấp và các điều kiện kinh tế bên ngoài khi đưa ra quyết định cấp vốn.

Đối với nhóm khách hàng có nguy cơ cao, ngân hàng nên thực hiện các biện pháp giám sát chặt chẽ như tăng lãi suất, hạn chế cấp vốn mới, và theo dõi sát sao dòng tiền cùng lịch sử giao dịch. Đồng thời, với những khách hàng tiềm năng nhưng có nguy cơ cao do các yếu tố khách quan, ngân hàng có thể triển khai các gói hỗ trợ tài chính, giãn nợ hoặc giảm lãi suất nhằm giúp họ vượt qua khó khăn tạm thời. Điều này không chỉ giảm nguy cơ nợ xấu mà còn tạo cơ hội phát triển lâu dài cho cả ngân hàng và khách hàng.

Việc tái cấu trúc nợ xấu là một trong những giải pháp quan trọng, bao gồm cơ cấu lại khoản vay, thỏa thuận giãn nợ hoặc giảm lãi suất, và chia nhỏ khoản trả nợ phù hợp với khả năng tài chính của khách hàng. Trong những trường hợp nợ xấu nghiêm trọng, việc áp dụng các biện pháp pháp lý để xử lý sẽ là cần thiết nhằm giảm thiểu tổn thất tài chính cho ngân hàng. Song song đó, việc triển khai các mô hình dự báo rủi ro sẽ giúp ngân hàng nhận diện sớm những khách hàng có nguy cơ chuyển từ nhóm nợ tốt sang nhóm nợ xấu. Sử dụng các kịch bản giả định dựa trên các yếu tố kinh tế vĩ mô như lạm phát hoặc tỷ lệ thất nghiệp cũng là cách hiệu quả để đánh giá rủi ro và chuẩn bị đối phó.

Ngân hàng cần xây dựng hệ thống cảnh báo sớm với các chỉ báo dựa trên lịch sử giao dịch, số dư tài khoản, và xu hướng thanh toán của khách hàng. Các biện pháp giám sát định kỳ như đánh giá lại hồ sơ tín dụng sẽ giúp phát hiện kịp thời những biến động và ngăn ngừa tình trạng nợ xấu lan rộng. Bên cạnh đó, việc hợp tác với Ngân hàng Nhà nước và các tổ chức quản lý tài chính khác để xử lý nợ xấu thông qua các gói hỗ trợ tái cơ cấu hoặc công ty mua bán nợ sẽ góp phần tăng cường hiệu quả quản lý.

Nâng cao ý thức tài chính của khách hàng cũng là một giải pháp quan trọng. Ngân hàng có thể triển khai các chương trình đào tạo quản lý tài chính cá nhân nhằm giúp khách hàng hiểu rõ hơn về các rủi ro tín dụng và cách quản lý nguồn vốn hiệu

quả. Đồng thời, việc khuyến khích lịch sử tín dụng tốt thông qua ưu đãi giảm lãi suất hoặc tăng hạn mức vay sẽ tạo động lực cho khách hàng duy trì khả năng thanh toán đúng hạn.

Cuối cùng, ngân hàng cần đa dạng hóa danh mục tín dụng để giảm thiểu rủi ro tập trung vào các ngành nghề rủi ro cao như bất động sản hay xây dựng. Hỗ trợ các doanh nghiệp nhỏ và vừa thông qua các gói tín dụng ưu đãi sẽ giúp ngân hàng mở rộng cơ hội phát triển và giảm áp lực từ nhóm khách hàng nợ xấu. Tất cả các giải pháp trên cần được triển khai đồng bộ để không chỉ xử lý các vấn đề nợ xấu hiện tại mà còn xây dựng nền tảng bền vững cho hoạt động tín dụng trong tương lai.

DANH MỤC THAM KHẢO

Altman, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Basel Committee on Banking Supervision. (2006). International convergence of capital measurement and capital standards: A revised framework. Bank for International Settlements. Retrieved from https://www.bis.org

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. AAAI Workshop on Learning for Text Categorization. Retrieved from https://www.cs.cmu.edu/~mccallum/papers/bayes-aaaiws98.pdf

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106. https://doi.org/10.1023/A:1022643204877

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408. https://doi.org/10.1037/h0042519

Saunders, A., & Cornett, M. M. (2018). Financial institutions management: A risk management approach (9th ed.). McGraw-Hill Education.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research, 1, 211–244. Retrieved from https://www.jmlr.org/papers/v1/tipping01a.html

Ngân hàng Nhà nước Việt Nam. (2021). Thông tư số 11/2021/TT-NHNN về phân loại nợ. Retrieved from https://www.sbv.gov.vn