TÊN ĐỀ TÀI:

PHÂN TÍCH DỮ LIỆU TÀI CHÍNH VÀ ĐỀ XUẤT MÔ HÌNH DỰ ĐOÁN RỦI RO VÀ CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN VAY TÀI CHÍNH: TRƯỜNG HỢP CỦA CÔNG TY TÀI CHÍNH HOME CREDIT

TP.HCM, ngày 15 tháng 6 năm 2023

Mục lục

TP.HCM, ngày 15 tháng 6 năm 2023

TP.HCM, ngày 15 tháng 6 năm 2023

TP.HCM, ngày 15 tháng 6 năm 2023

CHƯƠNG 1: TỔNG QUAN BÁO CÁO

1.1. Giới thiệu về công ty tài chính TNHH MTV Home Credit Việt Nam

Trong lĩnh vực tài chính, vay tiền là một phương thức quan trọng giúp cá nhân và doanh nghiệp đáp ứng nhu cầu tài chính và thúc đẩy sự phát triển kinh tế. Tuy nhiên, việc vay tài chính không đúng cách hoặc không được quản lý cẩn thận có thể mang đến những rủi ro và tác động tiêu cực đến cá nhân vay và hệ thống tài chính nói chung. Trong bối cảnh này, đặc biệt là trong lĩnh vực vay tín dụng cá nhân, việc đánh giá và quản lý rủi ro vay trở thành một yếu tố quan trọng để đảm bảo sự ổn định và bền vững cho hệ thống tài chính.

Công ty tài chính TNHH MTV Home Credit Việt Nam là một trong những công ty tài chính hàng đầu tại Việt Nam, với hơn 10 năm kinh nghiệm hoạt động trong lĩnh vực tài chính tiêu dùng. Công ty cung cấp các sản phẩm tài chính, bao gồm các khoản vay tiêu dùng và thẻ tín dụng, để đáp ứng nhu cầu của khách hàng về tiền mặt và thanh toán.

Theo báo cáo mới nhất của công ty (Home Credit Việt Nam, 2021), Home Credit đã và đang trở thành đối tác tài chính tin cậy của hàng triệu khách hàng trên toàn quốc. Công ty cam kết cung cấp các sản phẩm tài chính tiêu dùng chất lượng cao và đảm bảo trải nghiệm khách hàng tốt nhất thông qua dịch vụ khách hàng chuyên nghiệp và tiện lợi. Đồng thời, công ty cũng đầu tư vào công nghệ và đào tạo nhân viên để nâng cao chất lượng sản phẩm và dịch vụ.

Công ty tài chính TNHH MTV Home Credit Việt Nam là một đối tác tài chính tiêu dùng hàng đầu tại Việt Nam, đã và đang nghiên cứu và áp dụng các công nghệ tiên tiến để đảm bảo chất lượng sản phẩm và dịch vụ, cũng như quản lý rủi ro tín dụng một cách hiệu quả.

1.2. Đặt vấn đề

Trong ngành công nghiệp tài chính, Home Credit là một trong những công ty dẫn đầu về dịch vụ vay tín dụng cá nhân. Home Credit cung cấp các sản phẩm vay tiền nhanh chóng và thuận tiện cho các khách hàng có thu nhập thấp hoặc không có lịch sử

tín dụng. Tuy nhiên, đi kèm với lợi ích của việc cung cấp dịch vụ vay dễ dàng là sự xuất hiện của các rủi ro tiềm ẩn.

Các rủi ro trong vay tài chính Home Credit có thể bao gồm:

- Rủi ro tín dụng: Sự không trả nợ hoặc trả nợ chậm có thể xảy ra khi khách hàng không đáp ứng được các khoản vay.
- Rủi ro thị trường: Thay đổi trong điều kiện kinh tế và tài chính có thể ảnh hưởng đến khả năng trả nợ của khách hàng và giá trị tài sản thế chấp.
- Rủi ro hoạt động: Quản lý không hiệu quả, thiếu kiểm soát trong quá trình cấp vay và thu hồi nợ có thể dẫn đến mất cân đối tài chính và các vấn đề pháp lý.

Tác động tiêu cực của vấn đề rủi ro vay tài chính:

- Tác động tài chính: Nợ nần từ việc vay tín dụng có thể gây áp lực tài chính lên người vay và gia đình họ. Việc không thể hoàn trả kịp thời có thể dẫn đến lãi suất phạt, phí trễ hạn và đồng thời ảnh hưởng xấu đến điểm tín dụng của khách hàng trong tương lai.
- Vấn đề xã hội: Sự gia tăng của rủi ro vay tài chính Home Credit có thể gây ra những vấn đề xã hội như tăng cường nợ nần cá nhân và gia đình, khó khăn trong việc tiếp cận các sản phẩm và dịch vụ tài chính khác, cũng như sự mất cân bằng trong phân phối tài chính.

Trong bối cảnh tài chính tiêu dùng đang có xu hướng phát triển mạnh mẽ tại Việt Nam, công ty Home Credit Việt Nam đã nghiên cứu và xây dựng mô hình dự đoán rủi ro thu hồi nợ để quản lý rủi ro tín dụng và đảm bảo sự ổn định của hoạt động kinh doanh. Mô hình này được xây dựng trên cơ sở thu thập và phân tích các dữ liệu tài chính, dữ liệu khách hàng và các chỉ số kinh tế xã hội. Qua đó, công ty có thể đưa ra các quyết định về việc chấp nhận, từ chối hoặc điều chỉnh khoản vay để giảm thiểu rủi ro thu hồi nợ.

1.3. Bài toán

Trong phạm vi nghiên cứu này, nhóm tập trung vào bài toán chính trong việc phân tích dữ liệu hồ sơ khách hàng của Home Credit trong quá khứ và xây dựng mô hình dự đoán rủi ro thanh toán của khách hàng công ty Home Credit, cụ thể là: Liên

quan đến việc xây dựng "Mô hình dự đoán rủi ro thanh toán trong cho việc cho vay tài chính" có độ chính xác cao nhất. Mô hình này sẽ giúp cho công ty HomeCredit đưa ra dự đoán về khả năng thanh toán của khách hàng, các yếu tố ảnh hưởng đến rủi ro cho việc vay tài chính. Từ đó, đưa ra những quyết định thông minh và chính xác hơn trong việc cấp vay và tăng khả năng thu hồi nợ.

Với bài toán này, nhóm hy vọng sẽ giúp cho công ty Home Credit hoạt động hiệu quả hơn trong việc đánh giá hồ sơ cho vay một cách nhanh chóng qua model đã xây dựng, có độ chính xác cao dự đoán được nhiều hồ sơ rủi ro và tăng khả năng thu hồi nợ.

1.4. Mục tiêu

- Nắm được các kiến thức về các dữ liệu liên quan đến các hồ sơ vay, bao gồm các thông tin về khách hàng, hợp đồng và các chỉ tiêu khác liên quan đến tài chính, để hiểu quá trình đánh giá và dự đoán rủi ro của các hồ sơ nợ này.

- Áp dụng và phát triển kiến thức về phân tích dữ liệu và máy học để đưa ra mô hình đề xuất và khuyến nghị cho doanh nghiệp trong việc tối ưu hóa quy trình đánh giá và phân loại các hồ sơ vay, giảm thiểu rủi ro nợ, tăng cường hiệu quả giao dịch kinh doanh và cải thiện trải nghiệm khách hàng.

1.5. Phương pháp thực hiện

Các phương pháp thực hiện dự kiến bao gồm:

- Phân tích dữ liệu: Sử dụng dữ liệu lịch sử vay và thông tin khách hàng để phân tích mô hình rủi ro và nhận diện yếu tố ảnh hưởng đến khả năng trả nợ của khách hàng.
- Đánh giá rủi ro: Xác định các yếu tố rủi ro cụ thể và đo lường mức độ rủi ro liên quan đến các khoản vay.
- Xây dựng mô hình và đề xuất các biện pháp giảm thiểu và quản lý rủi ro.

TP.HCM, ngày 15 tháng 6 năm 2023

1.6. Dự kiến kết quả và ý nghĩa

Kết quả dự kiến sẽ đóng góp vào việc nâng cao hiệu quả quản lý rủi ro trong vay tài chính Home Credit. Việc áp dụng các biện pháp quản lý rủi ro được đề xuất có thể giúp giảm thiểu rủi ro tín dụng, tăng cường sự ổn định và bền vững của hệ thống tài chính, và tạo điều kiện thuận lợi hơn cho khách hàng khi vay tiền. Kết quả này cũng có ý nghĩa trong việc nâng cao nhận thức về quản lý rủi ro tài chính và thúc đẩy sự phát triển bền vững trong lĩnh vực vay tài chính.

1.7. Cấu trúc bài báo cáo

Bài báo cáo gồm 7 chương:

- Chương 1: Giới thiệu tổng quan về công ty tài chính TNHH MTV Home Credit Việt Nam, vấn đề đặt ra, bài toán nghiên cứu, mục tiêu mong muốn của đề tài và cấu trúc được áp dụng để thực hiện đề tài.
- Chương 2: Đưa ra nền tảng lý thuyết về tài chính, rủi ro tài chính, thẩm định khoản vay trong tài chính và các phương pháp khai phá dữ liệu như Dự đoán rủi ro nợ xấu, Phân loại khách hàng, XGBoost, Random Forest, K-Means.
- Chương 3: Trình bày phương pháp thực hiện quy trình và giới thiệu bộ dữ liệu mẫu.
- Chương 4: Trình bày phân tích khám phá dữ liệu (EDA) với 4 bảng: credit_card_balance, POS_CASH_balance, previous_application, installment_payments.
- Chương 5: Trình bày các thực nghiệm được thực hiện, bao gồm tiền xử lý dữ liệu (gồm Feature Engineering, làm sạch dữ liệu, mã hóa dữ liệu, lựa chọn đặc trưng, xử lý dữ liệu mất cân bằng), mô hình dự đoán rủi ro thanh toán (bao gồm 3 mô hình: Logistics regression, Random Forest, XGBoost) và mô hình phân loại hồ sơ đăng ký vay.
- Chương 6: Phần kết quả và thảo luận, trình bày kết quả đạt được và đề xuất với từng nhóm khách hàng.
- Chương 7: Phần kết luận, tổng kết lại các kết quả đạt được và đề xuất những hướng phát triển tiếp theo.

TP.HCM, ngày 15 tháng 6 năm 2023

Cấu trúc báo cáo được thiết kế một cách hợp lý và trung thực để trình bày các phần quan trọng của một báo cáo nghiên cứu, từ giới thiệu, lý thuyết, phương pháp, phân tích, thực nghiệm, kết quả, đề xuất và kết luận. Bằng cách này, báo cáo đảm bảo tính toàn diện và chính xác của các thông tin được trình bày, cũng như giúp người đọc có cái nhìn tổng quan và hiểu rõ về đề tài.

TP.HCM, ngày 15 tháng 6 năm 2023

CHƯƠNG 2: NỀN TẢNG LÝ THUYẾT

2.1. Nghiệp vụ

2.1.1. Tài chính

Nghiệp vụ tài chính là một trong những lĩnh vực quan trọng của kinh doanh và tài chính. Nó bao gồm các hoạt động liên quan đến quản lý tài chính, đầu tư và tài trợ để giúp doanh nghiệp hoạt động hiệu quả và đạt được lợi nhuận cao nhất. Tài chính là khía cạnh quan trọng của nghiệp vụ tài chính, bao gồm việc quản lý tiền, đầu tư và các hoạt động tài chính khác của doanh nghiệp để đảm bảo sự ổn định tài chính của nó (Chu et al., 2020).

2.1.2. Rủi ro tài chính

Trong lĩnh vực tài chính, quản lý rủi ro được coi là một yếu tố quan trọng để đảm bảo sự ổn định tài chính của doanh nghiệp. Quản lý rủi ro tài chính bao gồm đánh giá, kiểm soát và giảm thiểu tác động của rủi ro tài chính đến doanh nghiệp. Để đạt được mục tiêu này, các công ty tài chính cần đưa ra các giải pháp giảm thiểu rủi ro và theo dõi hiệu quả của chúng (Chu et al., 2020).

Rủi ro tín dụng là một trong những thách thức lớn nhất mà các công ty tài chính phải đối mặt. Đánh giá rủi ro tín dụng là quá trình đánh giá khả năng trả nợ của khách hàng và đánh giá rủi ro của khoản vay để đảm bảo rủi ro tín dụng thấp nhất. Quá trình đánh giá này bao gồm thu thập thông tin về khách hàng, đánh giá khả năng trả nợ của khách hàng và xác định rủi ro của khoản vay (Dinh et al., 2020).

2.1.3. Thẩm định khoản vay trong tài chính

Thẩm định khoản vay trong tài chính là quá trình đánh giá khách hàng và khoản vay của họ để đảm bảo rủi ro tín dụng thấp nhất. Quá trình này bao gồm đánh giá khả năng trả nợ của khách hàng và đánh giá rủi ro tín dụng của khoản vay (Chen, 2019). Đánh giá khả năng trả nợ của khách hàng bao gồm việc thu thập thông tin về thu nhập, lịch sử tín dụng và các khoản vay trước đó của khách hàng. Đánh giá rủi ro tín dụng của khoản vay bao gồm việc đánh giá tính khả thi của dự án, độ tin cậy của tài sản thế chấp và khả năng chi trả lãi và gốc của khoản vay.

Tóm lại, nền tảng lý thuyết trong nghiệp vụ tài chính là rất quan trọng để xây dựng mô hình dự đoán rủi ro thu hồi nợ cho các công ty tài chính. Các khái niệm về quản lý rủi ro tài chính, đánh giá rủi ro tín dụng và thẩm định khoản vay trong tài chính cũng là những yếu tố quan trọng cần được hiểu rõ để áp dụng cho thực tiễn.

2.2. Khai phá dữ liệu

2.2.1. Dự đoán rủi ro nợ xấu

Loan default risk prediction (Dự báo rủi ro nợ vay) là quá trình dự đoán khả năng một khoản vay sẽ gặp rủi ro nợ xấu, tức là người vay không thể trả nợ theo thỏa thuận ban đầu. Quá trình này sử dụng các phương pháp phân tích dữ liệu, mô hình dự đoán và thu thập thông tin về người vay để đưa ra dự báo về rủi ro nợ vay.

Các yếu tố quan trọng có thể được xem xét bao gồm lịch sử tín dụng của người vay, thông tin tài chính, hồ sơ tín dụng, tình hình công việc và thu nhập, và các yếu tố khác có thể ảnh hưởng đến khả năng trả nợ của người vay. (Wongnaa et al., 2013)

Quá trình dự báo rủi ro nợ vay giúp các tổ chức tài chính và ngân hàng đánh giá và quản lý rủi ro trong việc cấp vay, xác định mức độ rủi ro, đưa ra quyết định về việc duyệt vay, định mức lãi suất và xây dựng các chiến lược quản lý rủi ro phù hợp. Điều này giúp giảm thiểu rủi ro mất tiền, tăng cường an toàn và bảo đảm tính ổn định của hoạt động cho vay. (Muriithi et al., 2016)

2.2.2. Phân loại khách hàng

Phân đoạn khách hàng (Customer segmentation) là quá trình chia cơ sở khách hàng thành các nhóm hoặc phân đoạn riêng biệt dựa trên các đặc điểm, hành vi hoặc sở thích chung. Quá trình này bao gồm phân tích dữ liệu khách hàng để xác định các điểm chung và phân đoạn khách hàng thành các nhóm đồng nhất, cho phép doanh nghiệp tùy chỉnh chiến lược cung cấp sản phẩm và trải nghiệm khách hàng để đáp ứng tốt hơn nhu cầu và sở thích của từng phân đoạn. (Schneider & G.P. (2011); Sari et al., 2011)

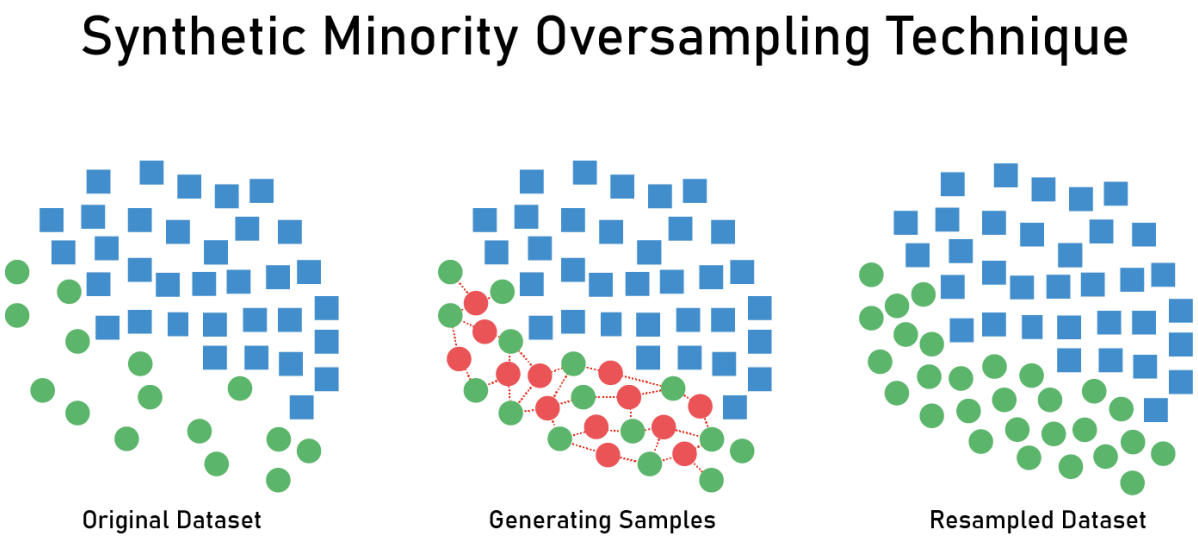
2.2.3. Dữ liệu mất cân bằng

Trong việc xử lý dữ liệu, tập dữ liệu không cân bằng là tình huống khi số lượng các mẫu thuộc một nhãn hoặc một lớp trong tập dữ liệu lớn hơn đáng kể so với số

lượng các mẫu thuộc những nhãn hoặc lớp khác. Điều này có thể gây ra hiện tượng mất cân bằng thông tin và có thể ảnh hưởng đến hiệu suất và độ chính xác của mô hình học máy.

Trong tập dữ liệu không cân bằng, mô hình học máy thường có xu hướng phân loại tập trung vào lớp đa số (nhãn/ lớp có số lượng mẫu nhiều hơn), trong khi lớp thiểu số (nhãn/lớp có số lượng mẫu ít hơn) được phân loại chưa chính xác hoặc bị bỏ qua. Điều này có thể gây ra các hệ quả không mong muốn trong việc phân loại hoặc dự đoán.

Một trong những giải pháp để xử lý tập dữ liệu không cân bằng là sử dụng các phương pháp tăng cường dữ liệu. SMOTE (Synthetic Minority Over-sampling Technique) là một trong số những thuật toán phổ biến được sử dụng để tăng cường dữ liệu trong trường hợp tập dữ liệu thiểu số. (Fernández,2018)



Hình 2.1.Thuật toán SMOTE cho dữ liệu mất cân bằng

(Nguồn: https://medium.com/analytics-vidhya/bank-data-smote-b5cb01a5e0a2)

SMOTE hoạt động bằng cách tạo ra các điểm dữ liệu tổng hợp dựa trên các điểm dữ liệu gốc của lớp thiểu số. Thay vì tạo ra các bản sao trùng lặp, SMOTE tạo ra các điểm dữ liệu tổng hợp có chút khác biệt so với các điểm dữ liệu gốc. Qua đó, SMOTE giúp cân bằng số lượng mẫu giữa các lớp và cung cấp đủ dữ liệu cho việc huấn luyện mô hình học máy.

Với việc áp dụng SMOTE, chúng ta có thể tạo ra các điểm dữ liệu tổng hợp từ lớp thiểu số, từ đó cân bằng tập dữ liệu và cải thiện hiệu suất của mô hình học máy trong việc phân loại các lớp không cân bằng.

Thuật toán SMOTE hoạt động như sau:

- Rút một mẫu ngẫu nhiên từ lớp thiểu số.
- Đối với các quan sát trong mẫu này, xác định k hàng xóm gần nhất.
- Sau đó, chọn một trong những hàng xóm đó và xác định vectơ giữa điểm dữ liệu hiện tại và hàng xóm đã chọn.
- Nhân vector này với một số ngẫu nhiên từ 0 đến 1.
- Để thu được điểm dữ liệu tổng hợp, cộng vectơ này vào điểm dữ liệu hiện tại.
- Thao tác này thực chất tương tự như việc di chuyển một chút điểm dữ liệu theo hướng của hàng xóm. Điều này giúp đảm bảo điểm dữ liệu tổng hợp không phải là một bản sao chính xác của một điểm dữ liệu đã tồn tại, đồng thời cũng đảm bảo rằng nó không quá khác biệt so với các quan sát đã biết trong lớp thiểu số.

2.2.4. Logistics regression

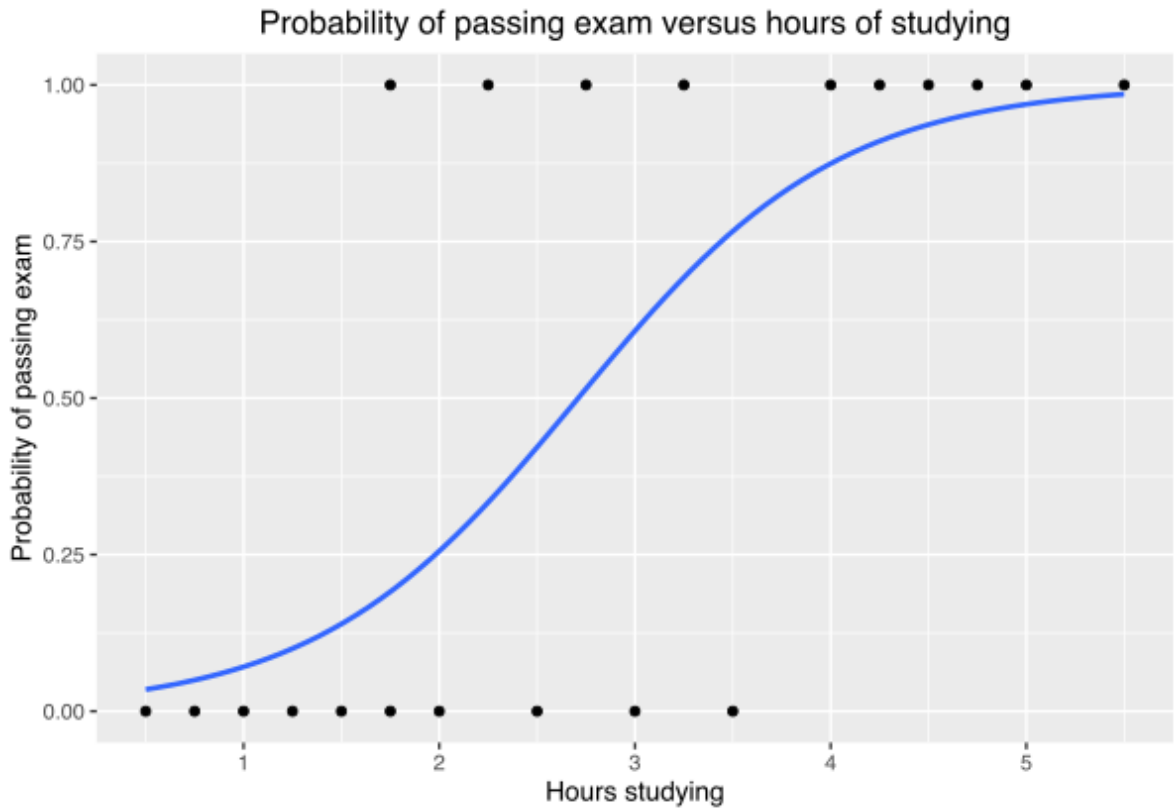
Hồi quy Logistic là một thuật toán phân loại được sử dụng để gán các đối tượng vào một tập hợp rời rạc các giá trị (như 0, 1, 2, ...). Một ví dụ điển hình là phân loại email, bao gồm email công việc, email cá nhân, email rác,... Cũng như xác định giao dịch trực tuyến có an toàn hay không an toàn, hoặc xác định các khối u lành tính hay ác tính,…

Công thức:

$$\widehat{y_i} = \sigma(w_0 + w_1 * x_1^{(i)} + w_2 * x_2^{(i)} = \frac{1}{1+e^{-(w_0+w_1*x_1^{(i)}+w_2*x_2^{(i)})}} \qquad (2.1)$$

Trong đó:

- yi là kết quả đầu ra, có giá trị từ 0 đến 1
- x1, x2,...,xi là các biến đầu vào, dựa trên tập dữ liệu mà chúng ta có
- w1, w2,...,wj là các hệ số của phương trình hồi quy
- σ là hàm sigmoid, được sử dụng để chuyển đổi tổng của các hệ số và biến đầu vào thành một giá trị giữa 0 và 1



Hình 2.2. Thuật toán hồi quy logictics

(Nguồn: https://en.wikipedia.org/wiki/Logistic_regression)

Nhiệm vụ của Hồi quy Logistic là tìm các hệ số của phương trình hồi quy dựa trên tập dữ liệu đầu vào của các biến độc lập và đầu ra là biến phụ thuộc. Điểm khác biệt của Hồi quy Logistic so với dự đoán liên tục như doanh thu, kích thước cơ thể,... là phương pháp này cung cấp 2 kết quả Đúng hoặc Sai (0,1) thay vì dự đoán một điều gì đó liên tục.

Mặc dù Hồi quy Logistic là một phương pháp truyền thống, nhưng theo nghiên cứu của (Christodoulou, 2019), không có bằng chứng rằng các mô hình học máy luôn hiệu quả hơn mô hình hồi quy tuyến tính. Hiệu quả của một mô hình học máy phụ thuộc vào nhiều yếu tố khác nhau.

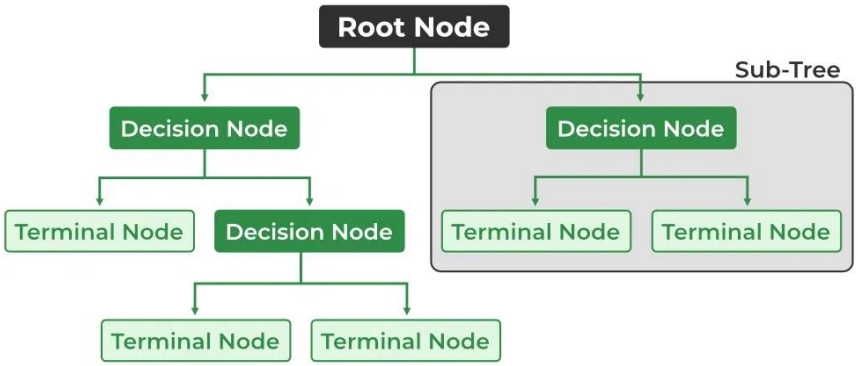
2.2.5. XGBoost

XGBoost, hay Extreme Gradient Boosting, là một thư viện gradient boosting phân tán tối ưu hóa. XGBoost cung cấp phương pháp parallel tree boosting như GBDT,

GBM và là học máy hàng đầu có thể được sử dụng để giải quyết được tất cả các vấn đề từ hồi quy (regression), phân loại (classification), ranking.

Thuật toán được xây dựng dựa trên học máy có giám sát (supervised machine learning), cây quyết định (decision trees), học kết hợp (ensemble learning) và gradient boosting (gradient boosting).

XGBoost sử dụng Cây quyết định làm mô hình cơ bản trong quá trình xây dựng. Mô hình Cây quyết định là một cấu trúc cây dạng sơ đồ, trong đó, nút gốc (Root Node) là nút cao nhất trong cây, đại diện cho toàn bộ tập dữ liệu, cũng sẽ là điểm bắt đầu của quá trình ra quyết định; mỗi nút Decision node đại diện cho một đặc trưng, các nhánh đại diện cho các quy tắc và các nút Terminal node đại diện cho kết quả của thuật toán. Cây quyết định sẽ đánh giá một cây câu hỏi true/false với mỗi đặc trưng theo nguyên tắc if-then-else để dự đoán nhãn.



Hình 2.3. Thuật toán Decision Tree

(Nguồn: https://www.geeksforgeeks.org/decision-tree/)

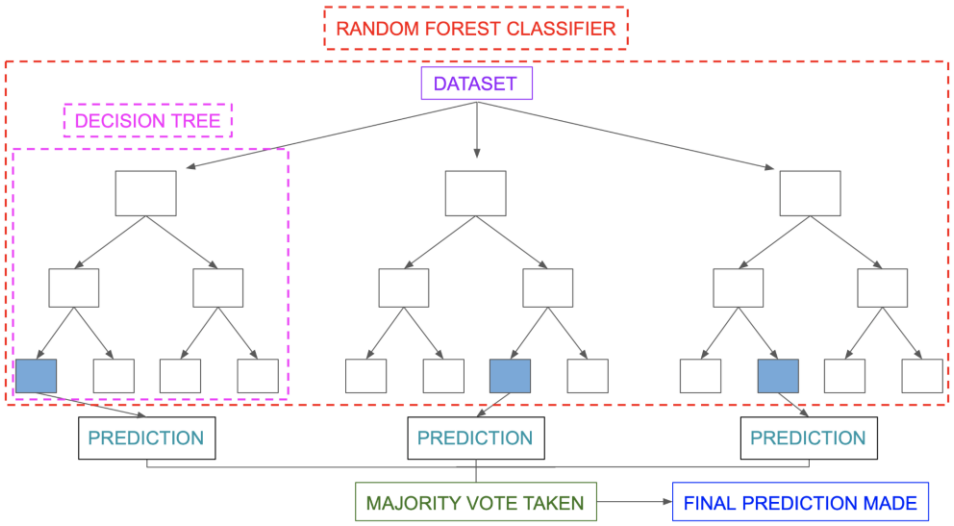
XGBoost cũng sử dụng học kết hợp để kết hợp kết quả từ tất cả các cây quyết định con và tối ưu quá trình dự đoán của các cây quyết định với phương pháp Gradient boosting.

XGBoost đã trở thành một trong những thuật toán học máy phổ biến và ứng dụng rộng rãi vì khả năng xử lý dữ liệu lớn, tăng tốc độ và hiệu suất huấn luyện (Chen et al., 2016).

2.2.6. Random Forest

Random Forest (Breiman, 2001) là một thuật toán học máy được sử dụng phổ biến, xây dựng nhiều cây quyết định bằng thuật toán Decision Tree. Đây là một thuật toán Học máy giám sát được sử dụng phổ biến trong các bài toán phân loại và hồi quy.

Random Forest sử dụng kỹ thuật Bagging, cụ thể xây dựng cùng lúc các cây quyết định từ các mẫu ngẫu nhiên khác nhau được lấy từ tập dữ liệu huấn luyện. Trước đó, thuật toán cần xác định ba tham số chính trước khi thực hiện huấn luyện mô hình, gồm kích thước của nút, số lượng cây và số lượng đặc trưng được lấy mẫu. Đây cũng là cách thức thuật toán Random Forest giải quyết các vấn đề liên quan đến hồi quy hoặc phân loại.



Hình 2.4. Thuật toán Random Forest xây dựng mô hình Cây quyết định với kỹ thuật Bagging

(Nguồn: https://bookdown.org/jcog196013/CancerMe/ranger.html)

Random Forest đem lại một số ưu điểm như loại bỏ hiện tượng quá khớp (overfitting), có khả năng xử lý các tập dữ liệu lớn, và cung cấp một ước lượng độ quan trọng của từng đặc trưng, cho phép đánh giá tầm quan trọng của từng đặc trưng trong mô hình. (IBM, n.d)

2.2.7. K-Means

K-Means là một phương pháp học máy không giám sát được sử dụng rộng rãi và nghiên cứu kỹ trong lĩnh vực khai phá dữ liệu, thường được áp dụng trong các bài toán phân cụm.

Ban đầu, thuật toán khởi tạo ngẫu nhiên một số lượng tâm cụm, là một điểm đại diện cho cụm và được tính bằng giá trị trung bình của tất cả các quan sát trong cụm đó.

Sau đó, thuật toán gán nhãn cho từng điểm dữ liệu dựa trên khoảng cách đến các tâm cụm và tiếp tục cập nhật lại vị trí của các tâm cụm. Với tập dữ liệu X (x_1, x_2,..., x_n), các điểm trung tâm cụm là C, thì khoảng cách Euclide giữa điểm dữ liệu và trọng tâm C là d(x,C) được tính với công thức sau:

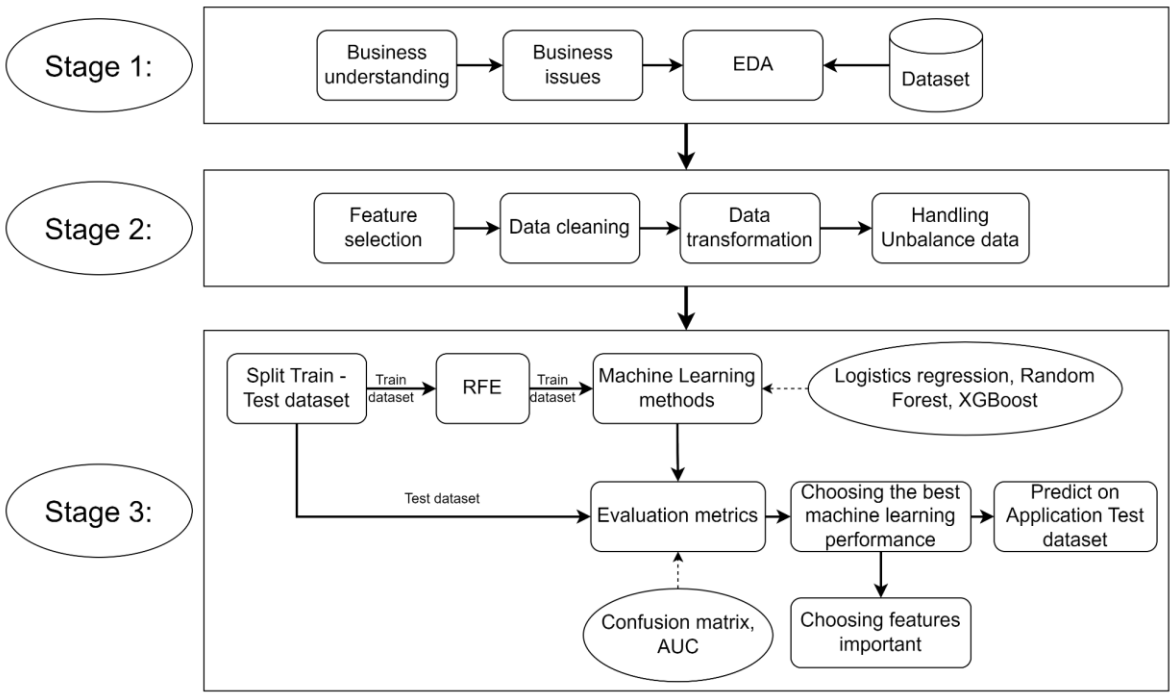
$$d(x, C_i) = \sqrt{\sum_{j=1}^{m} (x_j - C_{ij})^2} \qquad (2.2)$$

Quá trình này được lặp đi lặp lại cho đến khi thuật toán hội tụ và tìm ra một phân cụm tốt nhất cho dữ liệu.

Nhờ cách áp dụng đơn giản nhưng vẫn hiệu quả, K-Means được ứng dụng trong việc thực hiện phân cụm hay phân khúc khách hàng trong nhiều lĩnh vực khác nhau. (Jain et al., 2010)

CHƯƠNG 3: PHƯƠNG PHÁP THỰC HIỆN

3.1. Quy trình thực hiện



Hình 3.1. Mô hình nghiên cứu

(Nguồn: Nhóm tác giả)

Bài toán đánh giá khả năng thanh toán có thể được giải quyết thông qua quy trình được trình bày chi tiết trong Hình. Để thực hiện bài toán này, giai đoạn đầu tiên, nhóm tập trung vào việc hiểu về kinh doanh và nhận diện các vấn đề quan trọng liên quan đến rủi ro vay tài chính. Bộ dữ liệu Home Credit được sử dụng - một bộ sưu tập gần như đầy đủ các trường dữ liệu - để thực hiện phân tích tiền xử lý dữ liệu ban đầu (Exploratory Data Analysis - EDA). Giai đoạn này giúp hiểu cơ bản về công ty và nghiệp vụ tài chính Home Credit nắm bắt thông tin cơ bản về dữ liệu và hiểu rõ các vấn đề cần được giải quyết trong quy trình tiếp theo. Giai đoạn 2, tiến hành lựa chọn các đặc trưng quan trọng từ bộ dữ liệu để dự đoán rủi ro vay tài chính, xử lý dữ liệu các giá trị bị thiếu. Sau đó, chuyển đổi dữ liệu được thực hiện thành dạng phù hợp để áp dụng các phương pháp học máy. Ngoài ra, việc xử lý cân bằng dữ liệu không đồng đều để đảm bảo rằng mô hình học máy không bị thiên lệch do sự mất cân bằng của các lớp dữ liệu. Giai đoạn 3, tập dữ liệu được chia thành tập huấn luyện và tập kiểm tra để

đánh giá hiệu suất của các mô hình học máy. Chúng tôi sử dụng phương pháp Recursive Feature Elimination (RFE) để lựa chọn đặc trưng quan trọng. Sau đó, chúng tôi áp dụng các phương pháp học máy như Hồi quy Logistic, Random Forest và XGBoost để xây dựng mô hình. Các chỉ số đánh giá được sử dụng như ma trận nhầm lẫn và AUC để đánh giá hiệu suất của mô hình. Sau đó, mô hình học máy có hiệu suất tốt nhất được chọn và được dùng để lựa chọn các đặc trưng quan trọng ảnh hưởng đến phương pháp học máy đó. Cuối cùng, tập application test được dự đoán để đưa ra kết quả dự đoán về các hồ sơ có rủi ro vay tài chính.

3.2. Dataset

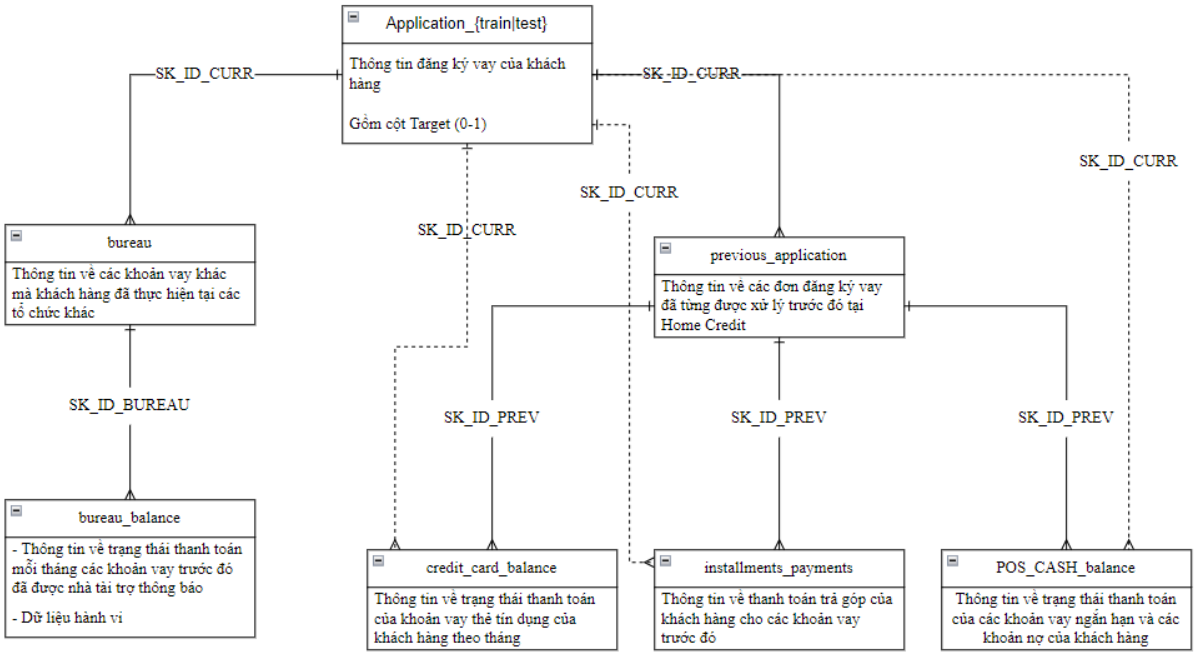
Nhóm sử dụng dataset Home Credit, được thu thập từ Kaggle, với 8 bảng dữ liệu với nội dung từng bảng được mô tả chi tiết qua Bảng:

Tên bảng	Nội dung
application_{train|test}.csv	Chứa thông tin về khách hàng đăng ký vay tiền.
POS_CASH_balance.csv	Chứa thông tin về trạng thái thanh toán của các khoản vay ngắn hạn và các khoản nợ của khách hàng.
bureau.csv	Chứa thông tin về các khoản vay khác mà khách hàng đã thực hiện tại các tổ chức khác.
bureau_balance.csv	Chứa thông tin về trạng thái thanh toán các khoản vay trước đó đã được nhà tài trợ thông báo.
credit_card_balance.csv	Chứa thông tin về trạng thái thanh toán của khoản vay thẻ tín dụng của khách hàng.
installments_payments.csv	Chứa thông tin về thanh toán trả góp của khách hàng cho các khoản vay trước đó.

previous_application.csv	Chứa thông tin về các đơn đăng ký vay đã từng được xử lý trước đó.

Bảng 3.1. Các bảng trong dataset Home Credit

Với sơ đồ thực thể được thể hiện như bên dưới:



Hình 3.2. Sơ đồ mối quan hệ thực thể (ERD)

(Nguồn: Nhóm tác giả)

Bảng	Mối quan hệ
application_{train|test} - previous_application	1-n
application_{train|test} - bureau	1-n
bureau - bureau_balance	1-n
previous_application - credit_card_balance	1-n
previous_application - POS_CASH_balance	1-n

TP.HCM, ngày 15 tháng 6 năm 2023

previous_application - installments_payments	1-n
application_{train|test} - credit_card_balance	1-n
application_{train|test} - POS_CASH_balance	1-n
application_{train|test} - installments_payments	1-n

Bảng 3.2. Mô tả mối quan hệ thực thể trong sơ đồ ER

CHƯƠNG 4: PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA)

4.1. Bảng credit_card_balance

4.1.1. Tổng quan về bảng

Bảng credit_card_balance mô tả số dư thẻ tín dụng hàng tháng của các thẻ tín dụng trước đây mà người đăng ký có với Home Credit được chụp lại dưới dạng bảng cân đối hàng tháng. Mỗi hàng trong bảng đại diện lịch sử của mỗi khoản tín dụng trước đó tại Home Credit (bao gồm tín dụng tiêu dùng và khoản vay tiền mặt) liên quan đến các khoản vay.

Bộ dữ liệu credit_card_balance có tổng số quan sát là 3840312, tức là bao gồm 3840312 hàng. Mỗi hàng trong bộ dữ liệu tương ứng với một giao dịch sử dụng thẻ tín dụng của khách hàng.

Bộ dữ liệu có tổng số cột là 23, tức là bao gồm 23 biến. Mỗi biến trong bộ dữ liệu mô tả các thuộc tính khác nhau của giao dịch thẻ tín dụng, chẳng hạn như số dư, khoản thanh toán tối thiểu, số tiền được thanh toán, tỷ lệ sử dụng tín dụng, v.v. Trong đó biến NAME_CONTRACT_STATUS trong bộ dữ liệu được xác định là biến phân loại (categorical value), 22 biến còn lại được xác định là biến số (numerical values).

4.1.2. Khám phá các biến

Thuộc tính	Mô tả	Ý nghĩa
SK_ID_PREV	ID của khoản vay trước đó trong hệ thống Home Credit.	Định danh khoản vay trước đó
SK_ID_CURR	ID của khoản vay trong tập dữ liệu mẫu	Định danh khoản vay hiện tại
MONTHS_BALANCE	Tháng cân đối số dư liên quan đến ngày nộp đơn vay	Là tháng cân bằng của số dư tín dụng của khách hàng, tính từ ngày nộp đơn xin tín dụng. Cột này

	(nếu giá trị là -1 thì có nghĩa là số dư mới nhất)	cho biết số tháng đã trôi qua kể từ thời điểm khách hàng gửi đơn xin tín dụng đến thời điểm cân bằng số dư tín dụng của khách hàng trong tháng đó.
AMT_BALANCE	Số dư trong tháng của khoản vay trước đó	Cho biết số tiền khách hàng đang nợ lại cho ngân hàng tại thời điểm cân bằng số dư tín dụng của tháng đó. Nếu khách hàng sử dụng một khoản tín dụng với ngân hàng và đã sử dụng một phần hoặc toàn bộ số tiền được cấp tín dụng, thì số dư tín dụng của khách hàng sẽ giảm dần theo thời gian và sẽ có số dư tín dụng còn lại tại thời điểm cân bằng số dư tín dụng trong tháng đó
AMT_CREDIT_LIMIT_ACTUAL	Hạn mức thẻ tín dụng trong tháng của khoản vay trước đó.	Cho biết số tiền tối đa mà khách hàng có thể sử dụng trong tháng đó. Nếu khách hàng đã được cấp một khoản tín dụng với giới hạn tín dụng là 50 triệu đồng và đã sử dụng một phần hoặc toàn bộ số tiền này, thì giới hạn tín dụng của khách hàng sẽ giảm dần theo thời gian và sẽ có giới hạn tín dụng mới tại thời điểm cân bằng số dư tín dụng trong tháng đó.
AMT_DRAWINGS_AT	Số tiền rút từ máy ATM trong tháng của khoản vay trước đó.	Nếu khách hàng sử dụng thẻ tín dụng để rút tiền mặt từ một máy ATM với số tiền là 1 triệu đồng, thì số dư tín dụng của khách hàng sẽ giảm đi 1

M_CURRENT		triệu đồng và giá trị của cột AMT_DRAWINGS_ATM_CURRENT sẽ được cập nhật với giá trị 1 triệu đồng.
AMT_DRAWINGS_CURRENT	Số tiền rút trong tháng của khoản vay trước đó.	
AMT_DRAWINGS_OTHER_CURRENT	Số tiền rút khác trong tháng của khoản vay trước đó.	
AMT_DRAWINGS_POS_CURRENT	Số tiền rút hoặc mua hàng trong tháng của khoản vay trước đó.	
AMT_INST_MIN_REGULARITY	Số tiền trả góp tối thiểu trong tháng của khoản vay trước đó.	
AMT_PAYMENT_CURRENT	Số tiền khách hàng trả trong tháng trên khoản vay trước đó.	
AMT_PAYMENT_TOTAL_CURRENT	Tổng số tiền khách hàng trả trong tháng trên khoản vay trước đó.	Các khoản thanh toán khác (ví dụ: thanh toán để tránh phạt, thanh toán trước hạn, v.v.).

AMT_RECEIVABLE_PRINCIPAL	Số tiền phải thu cho khoản gốc trên khoản vay trước đó.	Cho biết số tiền chính (gốc) mà khách hàng còn phải trả lại trên tín dụng của mình trong kỳ trước đó. Điều này có nghĩa là nếu khách hàng đã mượn một khoản tiền từ tín dụng và chưa trả hết khoản tiền đó, thì số tiền chính mà khách hàng còn nợ sẽ được cập nhật vào cột
AMT_RECIVABLE	Số tiền phải thu trên khoản vay trước đó	Cho biết tổng số tiền mà khách hàng còn phải trả lại trên tín dụng của mình trong kỳ trước đó, bao gồm cả số tiền chính (gốc) và số tiền lãi phải trả.
AMT_TOTAL_RECEIVABLE	Tổng số tiền phải thu trên khoản vay trước đó.	
CNT_DRAWINGS_ATM_CURRENT	Số lần rút tiền từ máy ATM trong tháng của khoản vay trước đó.	
CNT_DRAWINGS_CURRENT	Số lần rút tiền trong tháng của khoản vay trước đó.	
CNT_DRAWINGS_OTHER_CURRENT	Số lần rút tiền khác trong tháng của khoản vay trước đó.	
CNT_DRAWINGS_PO	Số lần rút tiền để mua hàng trong tháng của khoản vay	Số lần khách hàng đã thực hiện giao dịch rút tiền từ điểm bán hàng sử dụng thẻ tín dụng hoặc thẻ

S_CURRENT	trước đó.	ghi nợ nhiều nhất là 165 lần rút so với các nguồn khác như máy ATM. Trung bình cứ 10 người sẽ có 7 lần rút tiền trong tháng của khoản vay trước đó.
CNT_INSTALMENT_MATURE_CUM	cho biết số kỳ trả nợ đã được thanh toán trên tín dụng của khách hàng trong kỳ trước đó.	Nếu khách hàng đã có một khoản nợ tín dụng với 12 kỳ trả góp và đã thanh toán đầy đủ cho 6 kỳ trả góp trong kỳ trước đó của thẻ tín dụng, thì giá trị của cột CNT_INSTALMENT_MATURE_CUM sẽ là 6.
NAME_CONTRACT_STATUS	Tình trạng hợp đồng (đã ký, đang hoạt động, ...) của khoản vay trước đó.	"Active": Hợp đồng đang có hiệu lực và đang được thực thi. "Completed": Hợp đồng đã hoàn tất và không còn hiệu lực. "Signed": Hợp đồng đã được ký kết nhưng chưa có hiệu lực. "Demand": Hợp đồng đang trong trạng thái đòi nợ. "Sent proposal": Hợp đồng đã được gửi đề xuất nhưng chưa được ký kết. "Refused": Hợp đồng đã bị từ chối. "Approved": Hợp đồng đã được phê duyệt và đang chờ để được ký kết. Đưa ra các biện pháp để giải quyết nợ xấu khi hợp đồng nằm trong trạng thái "Demand", hoặc đưa ra các chiến lược để tối ưu hóa quá trình xử

		lý hợp đồng khi biết được trạng thái của hợp đồng.
SK_DPD	Số ngày mà một khoản vay hoặc tài khoản ngân hàng đã trễ thanh toán so với ngày đáo hạn.	Nếu khoản vay A có giá trị "SK_DPD" là 0, điều này cho thấy khoản vay này không có bất kỳ ngày trễ thanh toán nào. Nếu khoản vay B có giá trị "SK_DPD" là 10, có nghĩa là khoản vay này đã trễ thanh toán 10 ngày so với ngày đáo hạn.
SK_DPD_DEF	số ngày trễ thanh toán mặc định (Days Past Due) trong tháng với mức dung sai (tolerance) được áp dụng (các khoản nợ có số tiền vay thấp được bỏ qua) của khoản tín dụng trước đó	Trong một số trường hợp, các khoản nợ tín dụng có số tiền thấp (thường là nhỏ hơn một mức ngưỡng nhất định) có thể bỏ qua khi tính toán DPD để tránh những sai sót không đáng kể trong việc quản lý nợ. Nếu ngân hàng quy định rằng sự dung sai cho các khoản nợ tín dụng dưới 100.000 đồng là 5 ngày, và khách hàng có một khoản nợ tín dụng 60.000 đồng quá hạn 3 ngày, thì giá trị của cột SK_DPD_DEF sẽ là 3 ngày. Tuy nhiên, nếu khoản nợ tín dụng của khách hàng là 200.000 đồng và quá hạn 3 ngày, thì cả hai giá trị SK_DPD và SK_DPD_DEF đều là 3 ngày.

Bảng 4.1. Bảng khám phá và mô tả các biến trong bảng dữ liệu credit_card_balane

4.1.2.1. Kiểm tra và xem xét các giá trị rỗng

Các thuộc tính trong bảng credit_card_balance có các giá trị rỗng được trình bày trong hình dưới:

Cột “AMT_PAYMENT_CURRENT” (Số tiền khách hàng trả trong tháng trên khoản vay trước đó) trong tập dữ liệu chứa các giá trị bị rỗng (missing values) vì có

những tháng số tiền trả góp tối thiểu trong tháng của khoản vay trước đó bằng 0, trong trường hợp này, giá trị của cột “AMT_PAYMENT_CURRENT” sẽ bị thiếu, tức là không có số tiền nào được trả trong tháng đó.

Cột "AMT_DRAWINGS_ATM_CURRENT" trong tập dữ liệu có chứa các giá trị bị rỗng (missing values) vì có những tháng mà khách hàng không rút tiền mặt từ máy ATM trên khoản vay đó. Trong trường hợp này, giá trị của cột "AMT_DRAWINGS_ATM_CURRENT" sẽ bị thiếu, tức là không có số tiền nào được rút từ máy ATM trong tháng đó. Việc khách hàng sử dụng số tiền rút từ máy ATM để chi tiêu và thanh toán các khoản nợ trên khoản vay của mình, dẫn đến sự giảm số dư tín dụng.

Cột "CNT_DRAWINGS_POS_CURRENT" có thể bị thiếu giá trị rỗng (missing values) vì nó chỉ đếm số lần rút tiền để mua hàng trong tháng. Nếu khách hàng không thực hiện giao dịch mua hàng trong tháng đó, giá trị của cột sẽ bị thiếu.

Cột "AMT_DRAWINGS_POS_CURRENT" chứa các giá trị rỗng (missing values) vì khách hàng có thể không thực hiện giao dịch rút tiền hoặc mua hàng trong tháng đó.

Cột "CNT_DRAWINGS_OTHER_CURRENT" chứa các giá trị rỗng do khách hàng không thực hiện bất kỳ giao dịch rút tiền khác nào trong tháng trước đó. Có sự tương quan giữa 2 cột "CNT_DRAWINGS_OTHER_CURRENT" và "AMT_DRAWINGS_OTHER_CURRENT". Mỗi lần khách hàng rút tiền khác đều được ghi nhận trong cả hai cột. Nếu một giá trị trong cột "CNT_DRAWINGS_OTHER_CURRENT" bị thiếu, giá trị tương ứng trong cột "AMT_DRAWINGS_OTHER_CURRENT" cũng sẽ bị thiếu.

Cột "CNT_DRAWINGS_ATM_CURRENT" chứa các giá trị rỗng vì khách hàng có thể không thực hiện giao dịch rút tiền từ máy ATM trong tháng đó. Mức tương quan giữa cột "CNT_DRAWINGS_ATM_CURRENT" và "AMT_DRAWINGS_ATM_CURRENT" là 1 cho thấy sự tương quan hoàn hảo giữa số lần rút tiền từ máy ATM và số tiền rút từ máy ATM trên khoản vay của khách hàng. Điều này có nghĩa là mỗi lần khách hàng rút tiền từ máy ATM đều được ghi nhận trong

cả hai cột. Nếu một giá trị trong cột "CNT_DRAWINGS_ATM_CURRENT" bị thiếu, giá trị tương ứng trong cột "AMT_DRAWINGS_ATM_CURRENT" cũng sẽ bị thiếu.

Cột "AMT_INST_MIN_REGULARITY" có thể chứa các giá trị rỗng vì trên khoản vay của khách hàng, có thể không yêu cầu thanh toán tối thiểu trong mỗi kỳ hạn. Việc thanh toán số tiền tối thiểu phải được thực hiện dựa trên số dư trong tháng của khoản vay của khách hàng. Nếu giá trị trong cột "AMT_INST_MIN_REGULARITY" bị thiếu, khả năng cao giá trị tương ứng trong cột "AMT_BALANCE" cũng bị thiếu. (mức tương quan giữa 2 cột là 0.9)

	Total	**Percent**
AMT_PAYMENT_CURRENT	767988	19.998063
AMT_DRAWINGS_ATM_CURRENT	749816	19.524872
CNT_DRAWINGS_POS_CURRENT	749816	19.524872
AMT_DRAWINGS_OTHER_CURRENT	749816	19.524872
AMT_DRAWINGS_POS_CURRENT	749816	19.524872
CNT_DRAWINGS_OTHER_CURRENT	749816	19.524872
CNT_DRAWINGS_ATM_CURRENT	749816	19.524872
CNT_INSTALMENT_MATURE_CUM	305236	7.948208
AMT_INST_MIN_REGULARITY	305236	7.948208

Bảng 4.2. Bảng mô tả tỷ lệ missing data theo từng thuộc tính trong bảng "credit_card_balance"

4.1.2.2. Phân tích các biến số

a. Thống kê mô tả

	MONTHS_BALANCE	AMT_INST_MIN_REGULARITY	CNT_DRAWINGS_POS_CURRENT	CNT_INSTALMENT_MATURE_CUM	SK_DPD	SK_DPD_DEF
mean	-34.521921	3540.204129	0.559479	20.825084	9.283667	0.331622
std	26.667751	5600.154122	3.240649	20.051494	97.515700	21.479231
min	-96.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	-1.000000	202882.005000	165.000000	120.000000	3260.000000	3260.000000

Bảng 4.3. Bảng thống kê mô tả các biến trong bảng dữ liệu "credit_card_balance"

Thuộc tính MONTH_BALANCE:

Với giá trị tháng cập nhật thông tin tài khoản có giá trị min là -96 tháng, tức là thông tin về số dư tín dụng được cập nhật lâu nhất là 96 tháng trước thời điểm đăng ký, có thể đánh giá rằng tổ chức tín dụng không nhận được thông tin tài chính cập nhật thường xuyên từ khách hàng trong một khoảng thời gian đáng kể.

Trung bình, thời điểm cập nhật thông tin số dư tài khoản của 103,558 khách hàng tính từ tháng nộp đơn xin vay là xấp xỉ 34 tháng. Mặc dù giá trị này thấp hơn so với thời gian cập nhật lâu nhất, nhưng vẫn có một khoảng thời gian tương đối dài giữa thời điểm đăng ký và thời điểm cập nhật thông tin. Điều này có thể làm giảm hiệu quả trong việc quản lý rủi ro và thu hồi nợ, vì các thay đổi tài chính của khách hàng trong khoảng thời gian này không được phản ánh kịp thời.

Thuộc tính AMT_INST_MIN_REGULARITY:

Phân bố số tiền trả góp tối thiểu của 103,558 khách hàng có sự đa dạng lớn. Với mức trung bình là 3 triệu 540 nghìn VNĐ và mức cao nhất là 202 triệu 882 nghìn VNĐ, có thể thấy rằng có sự chênh lệch lớn giữa các khoản trả góp. Điều này có thể tạo ra rủi ro cho tổ chức tín dụng, vì một số khách hàng có thể gặp khó khăn trong việc trả nợ theo đúng lịch trình.

Đối với mức trả góp tối thiểu cao nhất là 202 triệu 882 nghìn VNĐ, có thể cho thấy một số khách hàng đang gánh nặng nợ quá lớn và gặp khó khăn trong việc trả nợ. Điều này tăng rủi ro cho tổ chức tín dụng vì khả năng thu hồi nợ từ những khoản vay lớn như vậy có thể gặp khó khăn.

Thuộc tính CNT_DRAWINGS_POS_CURRENT:

Số lần rút tiền từ điểm bán hàng sử dụng thẻ tín dụng hoặc thẻ ghi nợ nhiều nhất là 165 lần, so với các nguồn khác như máy ATM. Điều này có thể cho thấy một xu hướng rút tiền thường xuyên từ điểm bán hàng, có thể tạo ra một rủi ro tăng lên cho tổ chức tín dụng. Sự tăng cường giám sát và kiểm soát đối với các giao dịch rút tiền từ điểm bán hàng có thể là cần thiết để giảm thiểu rủi ro nợ tài chính và bảo vệ lợi ích của tổ chức tín dụng.

Trung bình, mỗi 10 khách hàng sẽ có khoảng 7 lần rút tiền trong một tháng cho khoản vay trước đó. Mức độ sử dụng tiền mặt trong khoản vay này có thể đồng nghĩa với việc khách hàng đang gặp khó khăn trong việc quản lý tài chính cá nhân. Điều này có thể làm tăng nguy cơ nợ quá hạn và khả năng thu hồi nợ hiệu quả.

Thuộc tính CNT_ INSTALMENT_ MATURE_ CUM:

Giá trị trung bình của cột CNT_INSTALMENT_MATURE_CUM là 20.825084, có nghĩa là trung bình mỗi khách hàng đã thanh toán khoảng 21 kỳ trả nợ trong tín dụng của họ trong kỳ trước đó. Điều này cho thấy rằng đa số khách hàng đang có xu hướng thanh toán đầy đủ hoặc gần đầy đủ số tiền nợ của mình.

Giá trị nhỏ nhất của cột CNT_INSTALMENT_MATURE_CUM là 0 cho thấy có một số khách hàng không ghi nhận kỳ trả nợ nào trong kỳ trước đó ⇒ khả năng khách hàng chưa bắt đầu trả nợ hoặc có sự trễ chậm trong việc thanh toán. Tuy nhiên, giá trị max là 120, cho thấy rằng vẫn có một số khách hàng chưa thanh toán nợ của họ trong một khoảng thời gian rất dài. Điều này có thể làm tăng rủi ro nợ tài chính cho tổ chức tín dụng.

Giá trị std của cột CNT_INSTALMENT_MATURE_CUM là 20.051494, cho thấy rằng phân bố của số kỳ trả nợ đã được thanh toán trên tín dụng của khách hàng khá đồng đều. Tuy nhiên, độ lệch chuẩn này cũng cho thấy rằng vẫn có một số khách hàng có thể gây khó khăn trong việc thu hồi nợ.

Thuộc tính SK_DPD:

Giá trị trung bình (mean) của "SK_DPD" là 9.283667, với độ lệch chuẩn (std) là 97.515700. Giá trị tối thiểu (min) là 0 và giá trị tối đa (max) là 3260.

Thông qua giá trị trung bình, có thể thấy rằng trung bình số ngày trễ thanh toán trong khoản vay là khoảng 9.28 ngày. Điều này cho thấy có một phần khách hàng thường xuyên trễ hạn trong việc thanh toán khoản vay.

Độ lệch chuẩn cao (97.515700) và giá trị tối đa lớn (3260) cho thấy sự biến động mạnh về số ngày trễ thanh toán giữa các khoản vay. Có sự khác biệt lớn về mức độ nợ quá hạn giữa các khách hàng.

Thuộc tính SK_DPD_DEF:

Giá trị trung bình gần với 0 (0.331622) cho thấy hầu hết các khoản vay không có nợ quá hạn trong tháng tính trên khoản tín dụng trước đó. Tuy nhiên, độ lệch chuẩn (21.479231) và giá trị tối đa (3260) cho thấy cũng có một số khoản vay gặp vấn đề nợ quá hạn lớn hơn, gây tăng rủi ro nợ tài chính.

Với giá trị trung bình gần với 0, có thể cho rằng hầu hết các khoản vay đều được thu hồi nợ đúng hạn trong tháng tính trên khoản tín dụng trước đó. Tuy nhiên, sự chênh lệch giữa giá trị tối đa (3260) và giá trị trung bình (0.331622) cho thấy cũng có một số khoản vay gặp khó khăn trong quá trình thu hồi nợ, có số ngày trễ thanh toán mặc định lớn.

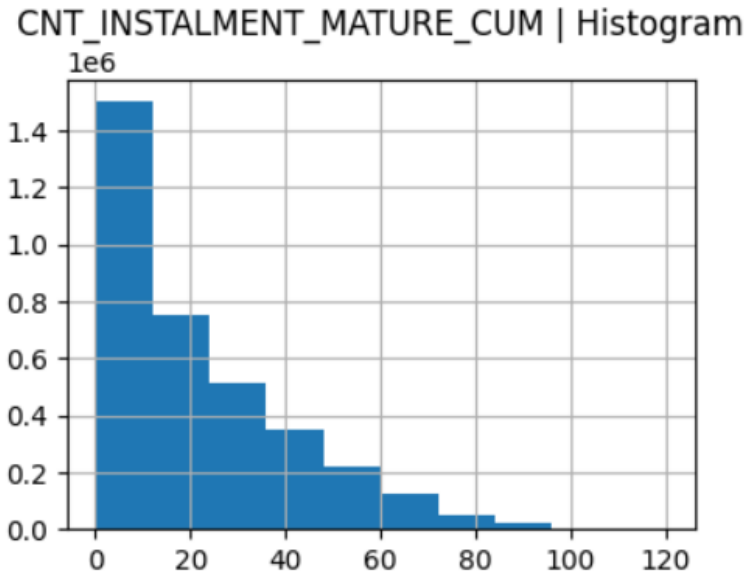
b. Phân tích phân phối

Cột "CNT_ INSTALMENT_ MATURE_ CUM":

Dựa vào histogram ta có thể thấy phân phối của cột "CNT_INSTALMENT_MATURE_CUM" có dạng lệch phải, với đỉnh của phân phối tập trung chủ yếu vào khoảng từ 1 đến 10 kỳ trả nợ đã được thanh toán trên tín dụng. Điều này cho thấy hầu hết khách hàng đã hoàn thành thanh toán cho một số lượng nhỏ kỳ trả nợ đầu tiên.

Tuy nhiên, khi xem xét các kỳ trả nợ tiếp theo, tỷ lệ thanh toán dần giảm. Điều này có thể cho thấy sự giảm tốc trong việc thanh toán nợ theo thời gian. Có thể rằng

khách hàng gặp khó khăn trong việc duy trì tốc độ thanh toán như ban đầu hoặc gặp các vấn đề tài chính khác.

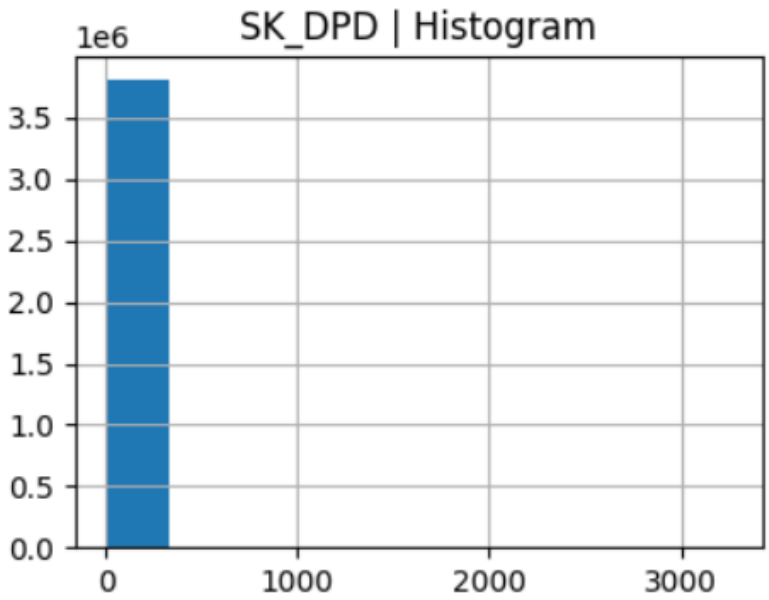


Hình 4.1. Biểu đồ phân phối của thuộc tính "CUM_ INSTALMENT_ MATURE_CUM"

(Nguồn: Nhóm tác giả)

Cột "SK_DPD":

Phân phối chủ yếu của cột "SK_DPD" tập trung vào khoảng thời gian 0 đến dưới 30 ngày quá hạn. Điều này cho thấy hầu hết các khoản vay trong bộ dữ liệu được xem xét đều không gặp vấn đề nợ quá hạn lớn trong tháng. Điều này có thể chỉ ra rằng người vay thường có xu hướng hoàn thành thanh toán đúng hạn hoặc chỉ chậm trễ trong khoảng thời gian ngắn.

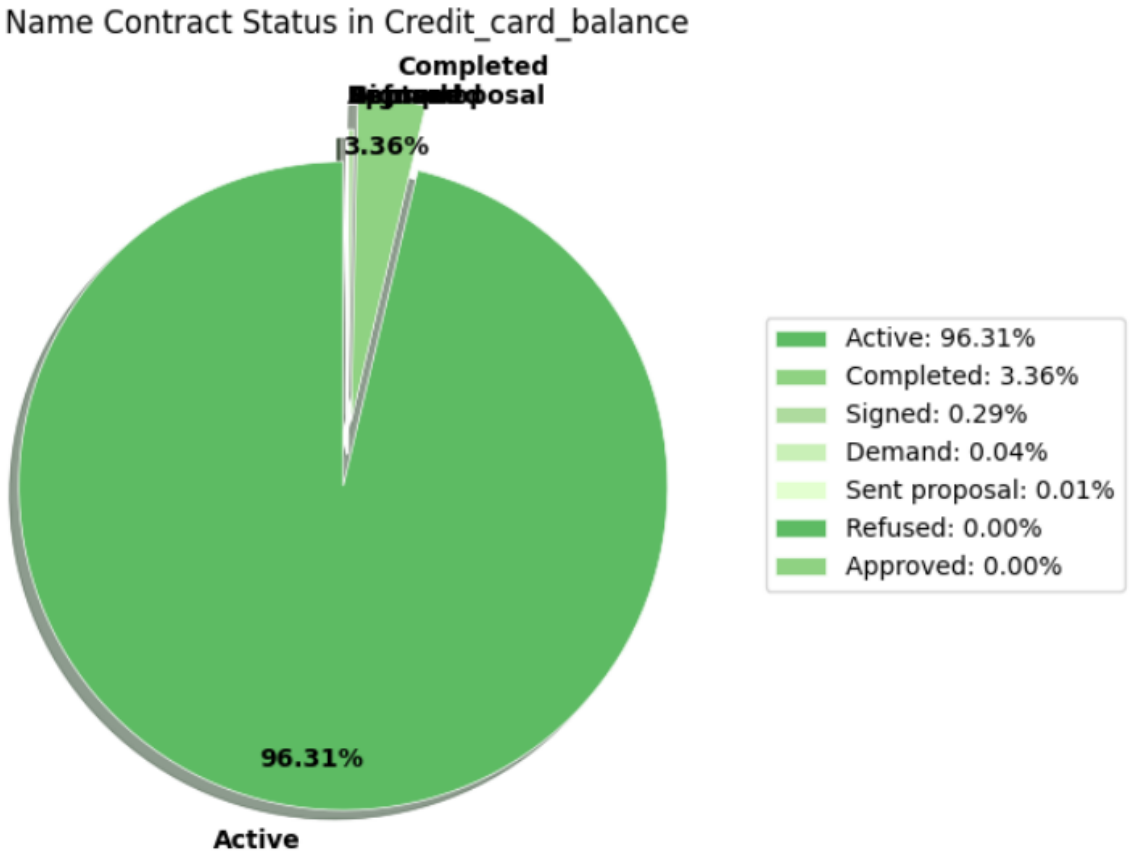


Hình 4.2. Biểu đồ phân phối của thuộc tính "SK_DPD"

(Nguồn: Nhóm tác giả)

Tuy nhiên, việc có một số khoản vay với số ngày quá hạn lớn hơn 30 ngày cũng có thể gây tăng rủi ro nợ tài chính. Cần chú ý đánh giá và quản lý kỹ lưỡng những khoản vay này để đảm bảo hiệu quả thu hồi nợ và giảm thiểu rủi ro tài chính cho tổ chức cho vay.

4.1.2.3. Phân tích biến phân loại



Hình 4.3. Biểu đồ thể hiện phần trăm các trạng thái của hợp đồng

(Nguồn: Nhóm tác giả)

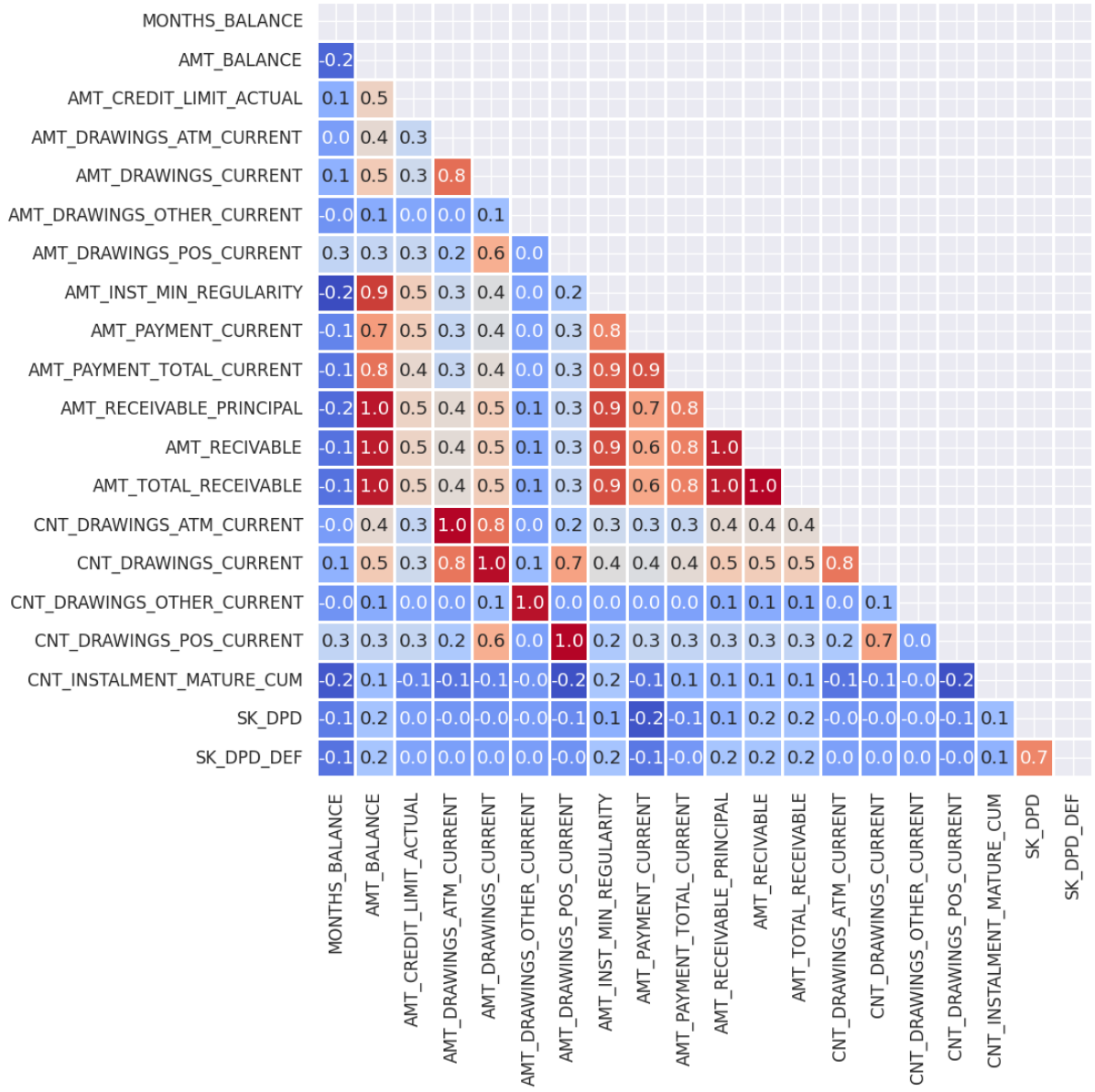
Biểu đồ phần trăm các trạng thái của hợp đồng cho thấy hầu hết các hợp đồng đều đang ở trạng thái active (được thực hiện và tiếp tục được thực hiện), chiếm tới 96.31%. Chỉ có một phần nhỏ các hợp đồng đã được hoàn thành (completed) hoặc đang ở các trạng thái khác như signed, demand, sent proposal

Thông qua phân tích biểu đồ, ta có thể phân nhóm khách hàng dựa trên trạng thái hợp đồng. Các khách hàng có hợp đồng ở trạng thái active có thể được coi là nhóm ổn định, trong khi các khách hàng có hợp đồng ở các trạng thái khác như signed, demand, sent proposal, refused hoặc approved có thể được coi là nhóm có rủi ro cao hơn.

Đối với các khách hàng có rủi ro cao hơn, đề xuất điều chỉnh lãi suất phù hợp có thể được áp dụng để đảm bảo tính ổn định của khoản vay và giảm thiểu rủi ro cho tổ chức cho vay.

4.1.3. Tương quan giữa các biến

4.1.3.1 Hệ số tương quan



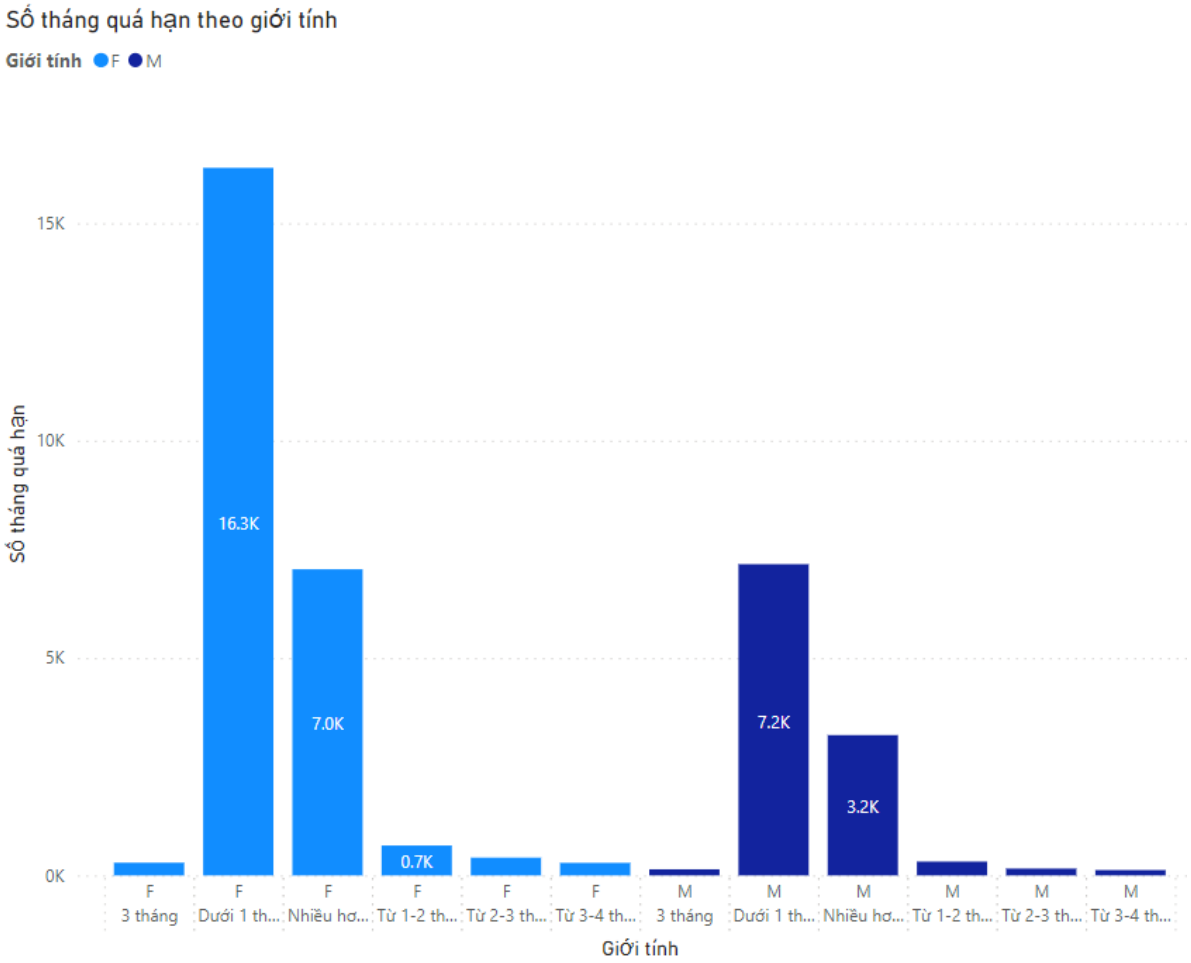
Hình 4.4. Biểu đồ thể hiện tương quan giữa các biến trong bảng "credit_card_balance"

(Nguồn: Nhóm tác giả)

TP.HCM, ngày 15 tháng 6 năm 2023

4.1.3.2 Trực quan hóa mối quan hệ giữa các biến

Phân tích số tháng khách hàng trả quá hạn theo giới tính:



Hình 4.5. Biểu đồ thể hiện số tháng khách hàng trả quá hạn theo giới tính

(Nguồn: Nhóm tác giả)

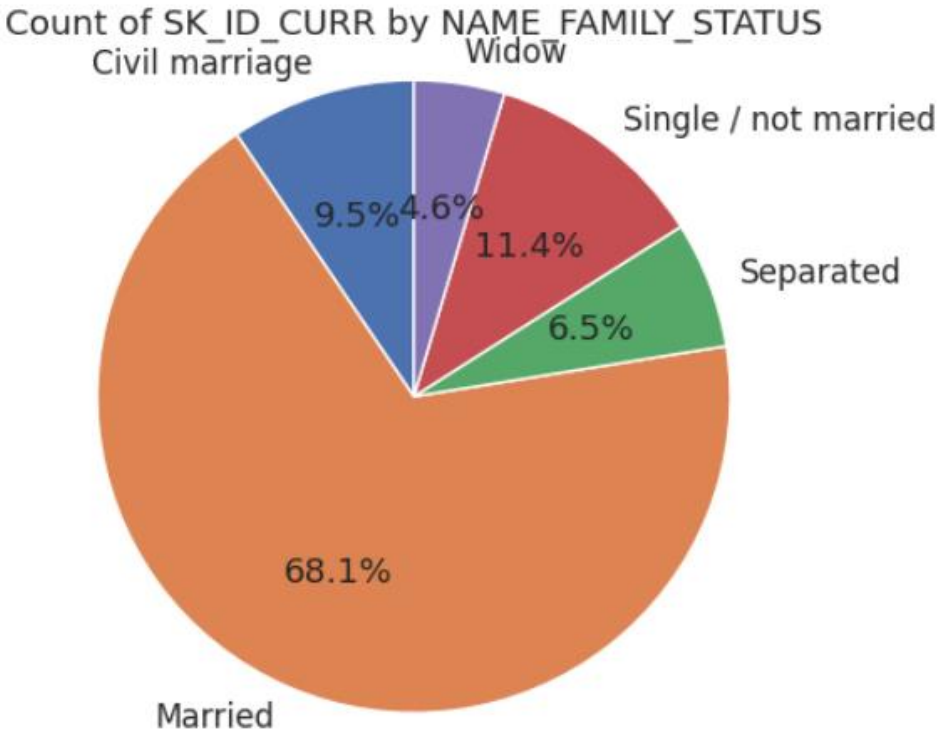
Dựa trên biểu đồ, ta có thể thấy rằng tỷ lệ khách hàng trễ hạn dưới 1 tháng là cao nhất cho cả giới tính Nam và Nữ. Tuy nhiên, tỷ lệ này lại cao hơn gấp đôi đối với khách hàng Nữ so với Nam. Điều này có thể cho thấy rằng khách hàng Nữ có xu hướng trễ hạn nhiều hơn so với khách hàng Nam.

Nếu xét đến khoảng thời gian trễ hạn lâu hơn nhiều, tỷ lệ khách hàng trễ hạn nhiều hơn 5 tháng của cả hai giới tính đều tương đối cao, tỷ lệ quá hạn của khách hàng Nữ cao gấp đôi khách hàng Nam.

Từ đó, có thể kết luận rằng việc quản lý nợ xấu cần được chú ý đối với cả hai giới tính, tuy nhiên, cần có một sự chú trọng đặc biệt đến khách hàng Nữ đối với

khoảng thời gian trễ hạn ngắn hơn 1 tháng. Các biện pháp cần được đưa ra để giảm thiểu tình trạng trễ hạn này, bao gồm việc cung cấp thông tin và hỗ trợ cho khách hàng Nữ để giúp họ quản lý tài chính và tránh việc nợ quá hạn.

Phân tích tỷ lệ khoản vay theo tình trạng gia đình:



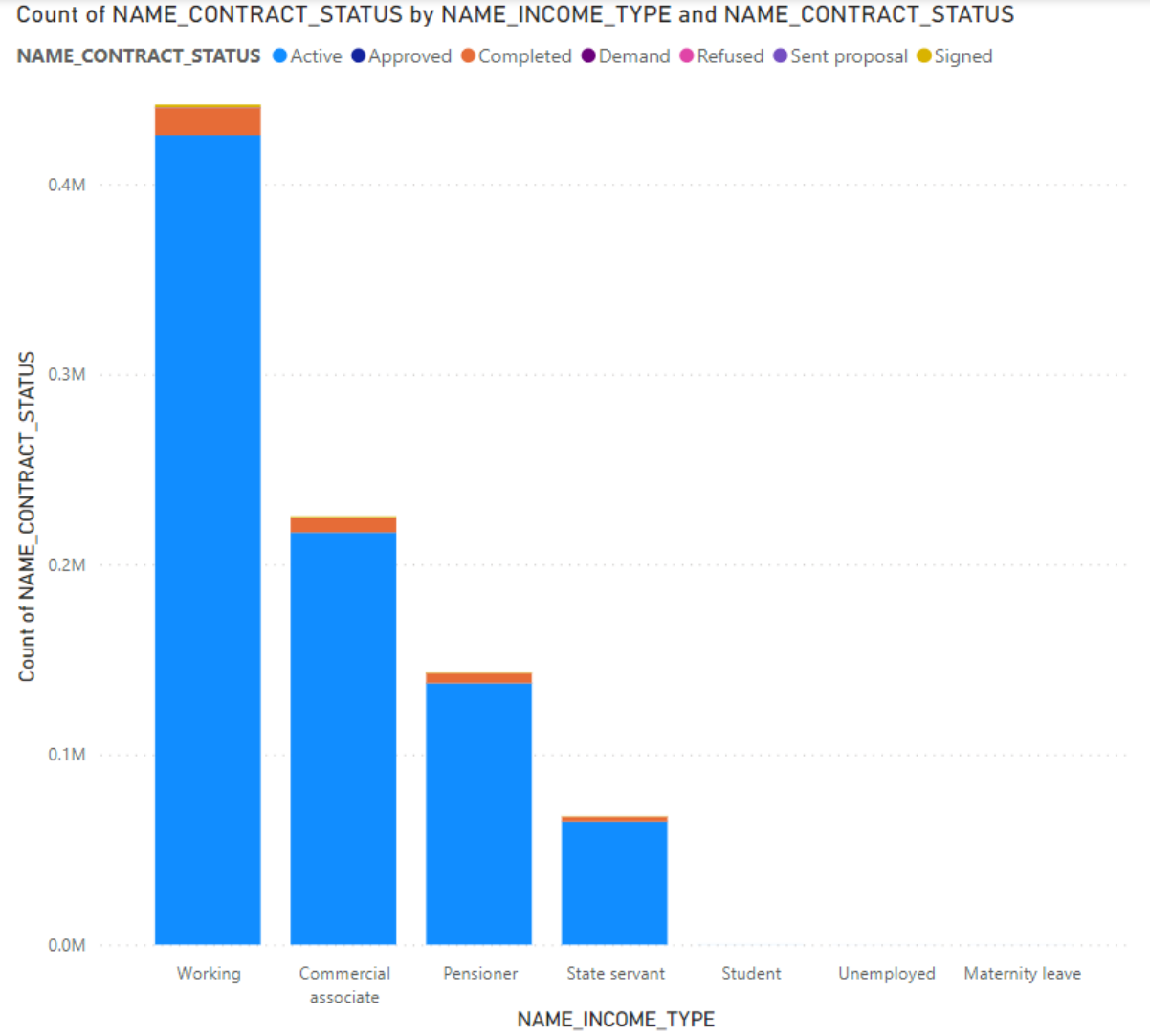
Hình 4.6. Biểu đồ thể hiện tỷ lệ khoản vay theo tình trạng gia đình

(Nguồn: Nhóm tác giả)

Dựa trên thông tin từ biểu đồ, ta có thể thấy rằng khách hàng đã kết hôn đóng góp nhiều đơn vay tín dụng nhất, chiếm tỷ lệ 68,1%. Điều này có thể cho thấy rằng khách hàng đã kết hôn cần phải đối mặt với nhiều thách thức tài chính hơn so với những người độc thân. Vì vậy họ có thể là một đối tượng khách hàng tiềm năng cho các chương trình cho vay tài chính.

Tuy nhiên, cần phải cân nhắc kỹ lưỡng việc cho vay tài chính cho khách hàng đã kết hôn và đảm bảo rằng họ có khả năng trả nợ đầy đủ và đúng hạn. Việc cho vay quá nhiều có thể dẫn đến tình trạng nợ xấu và ảnh hưởng đến tài chính cá nhân của khách hàng.

Phân tích tình trạng hợp đồng theo loại thu nhập:



Hình 4.7. Biểu đồ thể hiện tình trạng hợp đồng theo loại thu nhập

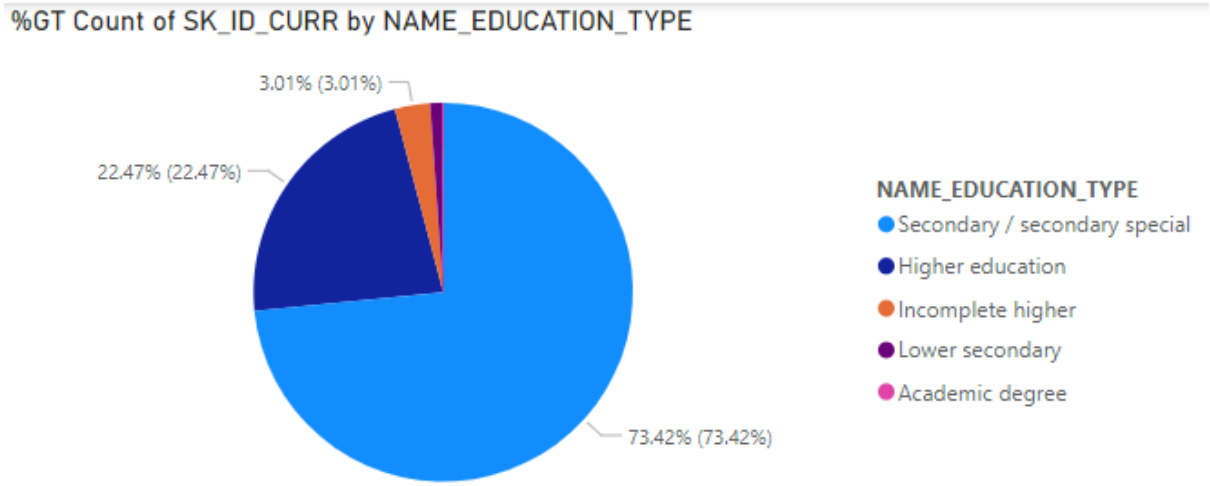
(Nguồn: Nhóm tác giả)

Dựa trên thông tin từ biểu đồ, ta có thể thấy rằng khách hàng có tình trạng thu nhập là "Working" có xu hướng có nhu cầu lớn về các khoản vay tài chính, và các hợp đồng vay của họ thường được phê duyệt nhiều hơn so với các nhóm khách hàng khác. Giá trị "working" trong cột này có nghĩa là khách hàng đang làm việc trong một công việc chính thức hoặc có thu nhập từ việc làm tự do.

Đây là một trong những loại thu nhập phổ biến và đa dạng nhất, thường được các tổ chức tín dụng đánh giá cao và có thể dễ dàng được chấp nhận cho các khoản vay tài chính.

Đứng sau đó là khách hàng làm việc trong lĩnh vực kinh doanh, thương mại hoặc có liên quan đến các công ty, tổ chức hoặc doanh nghiệp.

Phân tích trình độ học vấn của khách hàng được phê duyệt khoản vay:



Hình 4.8. Biểu đồ thể hiện trình độ học vấn của khách hàng được phê duyệt khoản vay

(Nguồn: Nhóm tác giả)

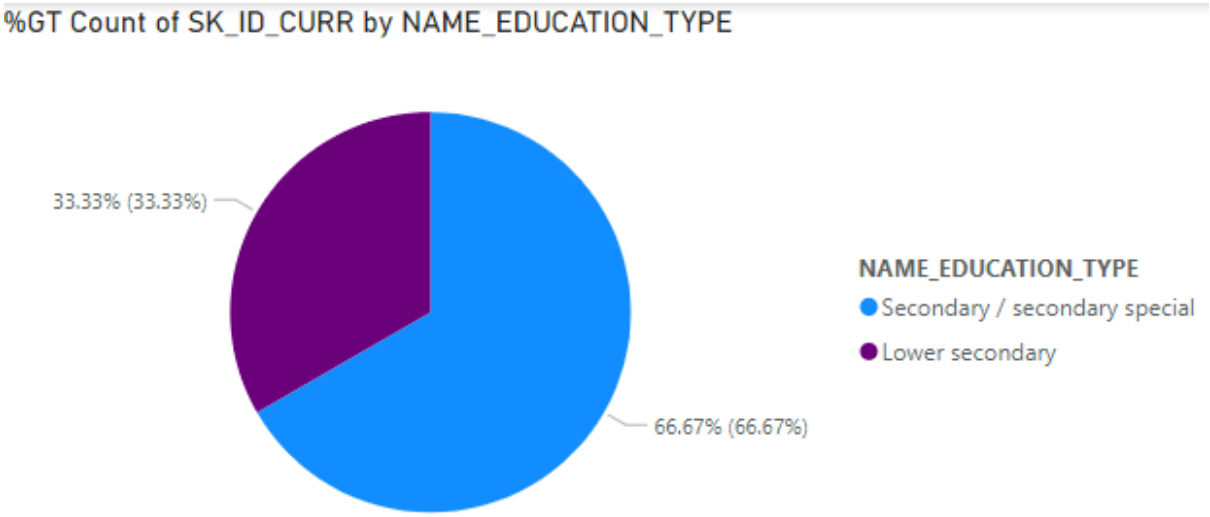
Dựa trên thông tin từ biểu đồ, ta có thể thấy rằng người nộp đơn có trình độ học vấn "Secondary/secondary special" (học vấn trung học phổ thông và chuyên nghiệp) được phê duyệt khoản vay chiếm tỷ lệ lớn nhất, với tỷ lệ là 73,42%. Trình độ học vấn "Higher education" (đại học và cao học) đứng thứ hai với tỷ lệ 22,4%. Người nộp đơn với trình độ học vấn "Incomplete higher" (học vấn chưa hoàn thành), "Lower secondary" (học vấn trung học cơ sở) và "Academic degree" (học vấn cao cấp) chiếm tỷ lệ rất ít.

Quyết định cho vay tài chính có xu hướng ưu tiên đối tượng có trình độ học vấn "Secondary/secondary special" và "Higher education". Điều này có thể cho thấy các trình độ học vấn này được coi là có khả năng trả nợ tốt hơn và có thể đáng tin cậy hơn trong việc quản lý và sử dụng tài chính.

Đối với những trình độ học vấn khác như "Incomplete higher", "Lower secondary" và "Academic degree", tỷ lệ đạt được khoản vay và khả năng thu hồi nợ có

thể thấp hơn. Cần có sự xem xét cẩn thận khi đánh giá khả năng trả nợ và xác định các biện pháp quản lý rủi ro cho những đối tượng này.

Phân tích trình độ học vấn của khách hàng bị từ chối khoản vay:



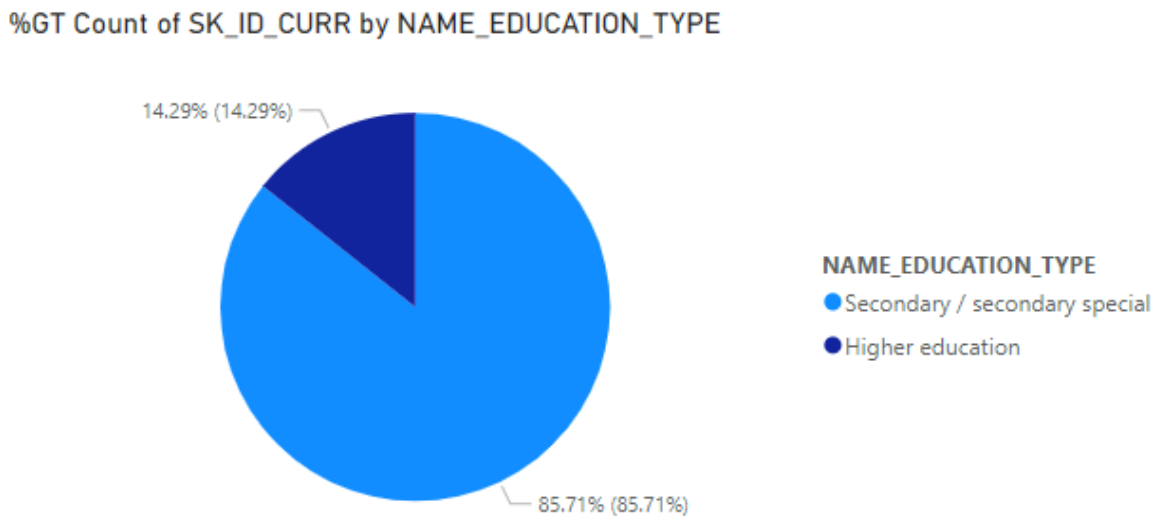
Hình 4.9. Biểu đồ thể hiện trình độ học vấn của khách hàng bị từ chối khoản vay

(Nguồn: Nhóm tác giả)

Biểu đồ trên cung cấp thông tin về trình độ học vấn của khách hàng bị từ chối khoản vay. Có hai giá trị được thể hiện trên biểu đồ: Secondary/secondary special và Lower secondary. Biểu đồ cho thấy rằng hầu hết khách hàng bị từ chối khoản vay có trình độ học vấn ở mức Secondary/secondary special (chiếm 66.67%), còn mức Lower secondary chiếm 33.33%.

Từ đó cho thấy khách hàng có trình độ học vấn thấp thường có khả năng thu nhập thấp và ít có kinh nghiệm trong quản lý tài chính cá nhân. Do đó, khách hàng này có thể gặp khó khăn trong việc thanh toán nợ và có nguy cơ cao về việc không trả nợ đúng hạn hoặc không trả nợ được. Vì vậy, số lượng khách hàng bị từ chối khoản vay với trình độ học vấn thấp là một yếu tố đáng quan ngại trong việc đánh giá rủi ro thu hồi nợ của một tổ chức tín dụng.

Phân tích trình độ học vấn của khách hàng đang trong trạng thái đòi nợ



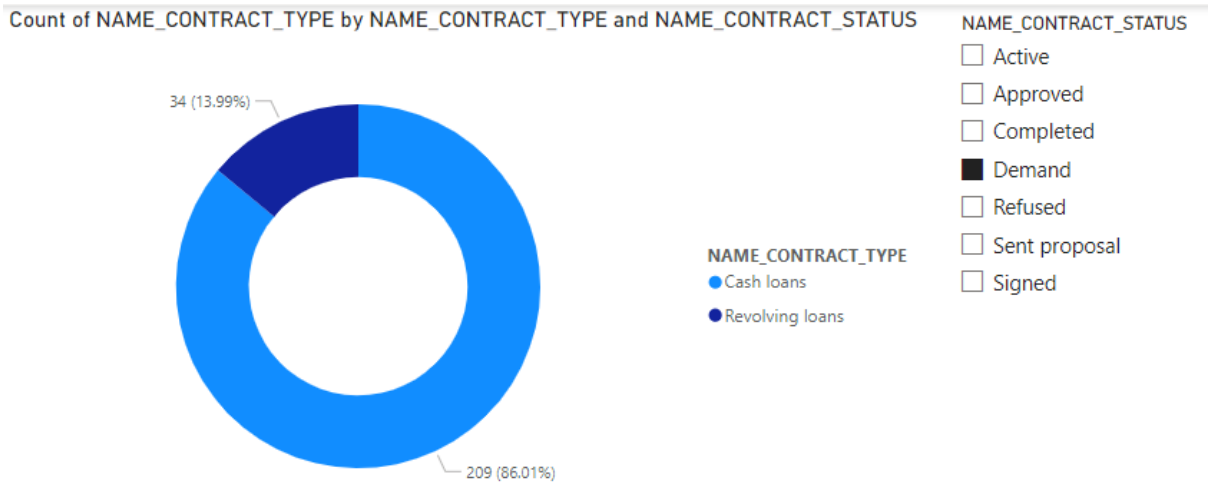
Hình 4.10. Biểu đồ thể hiện trình độ học vấn của khách hàng đang trong trạng thái đòi nợ

(Nguồn: Nhóm tác giả)

Thông qua biểu đồ trình độ học vấn của khách hàng đang trong trạng thái đòi nợ, ta có thể nhận thấy rằng hầu hết các khách hàng đang trong trạng thái đòi nợ đều có trình độ học vấn ở mức Secondary/secondary special (chiếm 85.71%) và chỉ có một số ít khách hàng có trình độ học vấn ở mức Lower secondary (chiếm 14.29%). Điều này cho thấy rằng tỷ lệ khách hàng đang trong trạng thái đòi nợ ở nhóm khách hàng có trình độ học vấn thấp là rất cao, và đây là một yếu tố đáng quan ngại trong việc đánh giá rủi ro thu hồi nợ của một tổ chức tín dụng.

Những khách hàng có trình độ học vấn thấp có thể có thu nhập thấp và ít kinh nghiệm trong việc quản lý tài chính cá nhân và cũng có thể không hiểu rõ về các điều khoản và điều kiện trong hợp đồng vay, dẫn đến nguy cơ không thể trả nợ đúng hạn hoặc không trả nợ được. Vì vậy, việc đánh giá rủi ro thu hồi nợ của một tổ chức tín dụng cần phải xem xét đến trình độ học vấn của khách hàng, đặc biệt là nhóm khách hàng có trình độ học vấn thấp.

Phân tích loại hợp đồng vay của các hồ sơ đang trong trạng thái đòi nợ nhiều nhất:



Hình 4.11. Biểu đồ thể hiện loại hợp đồng vay của các hồ sơ đang trong trạng thái đòi nợ nhiều nhất

(Nguồn: Nhóm tác giả)

Từ biểu đồ trên, chúng ta thấy rằng khoản vay là tiền mặt chiếm tỷ lệ lớn hơn trong trạng thái đòi nợ so với khoản vay là vay vòng. Tỷ lệ này là 86,01%, trong khi khoản vay là vay vòng chỉ chiếm 13,99%.

Điều này có thể liên quan đến việc khoản vay tiền mặt thường có lãi suất cao hơn, hoặc khách hàng có xu hướng sử dụng khoản vay tiền mặt để giải quyết các vấn đề tài chính khẩn cấp, vì vậy họ có thể gặp khó khăn hơn trong việc trả nợ.

4.1.4. Kết luận

Cột SK_DPD có ảnh hưởng lớn đến việc đánh giá rủi ro và hiệu quả thu hồi nợ. Số ngày quá hạn trong tháng của khoản vay trước đó càng cao. Điều đó cho thấy việc cho khách hàng này vay có rủi ro cao hơn so với các khách hàng có SK_DPD thấp. Việc đánh giá này còn cần nhiều yếu tố khác như: nhân khẩu học của khách hàng, số tiền mà khách hàng vay so với mức thu nhập hiện tại và hoàn cảnh hiện tại (số con cái, loại nhà ở, loại nghề nghiệp...).

Khoản vay là tiền mặt có khả năng gặp rủi ro nợ xấu cao hơn so với khoản vay là vay vòng, do đó các nhà cho vay cần đánh giá kỹ hơn trước khi cho vay khoản vay là tiền mặt.

Việc quản lý nợ xấu cần được chú ý đối với cả hai giới tính, tuy nhiên, cần có một sự chú trọng đặc biệt đến khách hàng Nữ đối với khoảng thời gian trễ hạn ngắn

hơn 1 tháng. Các biện pháp cần được đưa ra để giảm thiểu tình trạng trễ hạn này, bao gồm việc cung cấp thông tin và hỗ trợ cho khách hàng Nữ để giúp họ quản lý tài chính và tránh việc nợ quá hạn.

Khách hàng đã kết hôn có thể là một đối tượng khách hàng tiềm năng cho các chương trình cho vay tài chính. Tuy nhiên, cần phải cân nhắc kỹ lưỡng việc cho vay tài chính cho khách hàng đã kết hôn và đảm bảo rằng họ có khả năng trả nợ đầy đủ và đúng hạn.

Trình độ học vấn "Secondary/secondary special" và "Higher education" được coi là có khả năng trả nợ tốt hơn và có thể đáng tin cậy hơn trong việc quản lý và sử dụng tài chính. (NAME_CONTRACT_STATUS của 2 trình độ trên thể hiện khoản vay được phê duyệt, có giá trị: “active”, “approved”, “completed”, “sent proposal”, “signed”.)

Khoản vay tiền mặt là loại hợp đồng vay có hồ sơ đang trong trạng thái đòi nợ nhiều nhất. Điều này cho thấy khoản vay này thường có lãi suất cao hơn, hoặc khách hàng có xu hướng sử dụng khoản vay tiền mặt để giải quyết các vấn đề tài chính khẩn cấp, vì vậy họ có thể gặp khó khăn hơn trong việc trả nợ.

4.2. Bảng POS_CASH_balance

4.2.1. Tổng quan về bảng

Bảng "POS_CASH_balance" trong bộ dữ liệu Home Credit mô tả thông tin về các khoản vay POS (điểm bán hàng) và tiền mặt trước đó mà người đăng ký đã có với Home Credit. Đây là một bảng dữ liệu quan trọng để hiểu và theo dõi các giao dịch tiền mặt và vay tại các điểm bán hàng.

Bảng "POS_CASH_balance" chứa các thông tin như số ngày quá hạn thanh toán, số tiền trả góp hàng tháng, số tiền rút, số tiền hoàn lại và các thông tin khác liên quan đến các giao dịch tiền mặt và khoản vay POS. Bảng này giúp cung cấp cái nhìn tổng quan về tình trạng thanh toán và sử dụng tiền mặt của khách hàng tại các điểm bán hàng. Thông qua bảng dữ liệu, ta có thể đánh giá và phân tích sự tương tác giữa khách hàng và các điểm bán hàng, mức độ nợ quá hạn và tình trạng thanh toán của khách hàng. Điều này có thể giúp tổ chức cho vay đánh giá rủi ro tài chính và thu hồi nợ hiệu

quả, cũng như đưa ra các quyết định về việc cung cấp tín dụng cho khách hàng trong tương lai.

Bảng dữ liệu có tổng cộng 10,001,358 quan sát và 8 biến. Mỗi biến trong bộ dữ liệu mô tả các thuộc tính khác nhau của trạng thái và lịch sử thanh toán của khoản vay POS và tiền mặt trước đó của khách hàng. Trong đó biến NAME_CONTRACT_STATUS trong bộ dữ liệu được xác định là biến phân loại (categorical value), 7 biến còn lại được xác định là biến số (numerical values).

4.2.2. Khám phá các biến

Thuộc tính	**Mô tả**	**Ý nghĩa**
SK_ID_PREV	ID của khoản vay trước ở Home Credit liên quan đến khoản vay trong mẫu (Một khoản vay trong mẫu có thể có 0, 1, 2 hoặc nhiều khoản vay trước ở Home Credit)	Định danh khoản vay trước đó
SK_ID_CURR	ID của khoản vay trong hiện tại	Định danh khoản vay hiện tại
MONTHS_BALANCE	Tháng cập nhật thông tin về số dư tài khoản của khách hàng, tính từ tháng nộp đơn xin vay.	Đánh giá sự thay đổi trong số dư tài khoản của khách hàng trong một khoảng thời gian nhất định. Cho biết khoảng thời gian mà khách hàng đã sử dụng các tài khoản và có số dư trong tài khoản. Nếu giá trị của cột này lớn, tức là khách hàng

		đã sử dụng các tài khoản trong một khoảng thời gian dài và có số dư trong tài khoản, điều này có thể cho thấy khách hàng có khả năng quản lý tài chính tốt hơn và có thể có khả năng trả nợ đúng hạn. Tuy nhiên, nếu giá trị của cột này quá nhỏ, tức là khách hàng đã sử dụng các tài khoản trong một khoảng thời gian ngắn, điều này có thể cho thấy khách hàng có khả năng không quản lý tài chính tốt và có thể có khả năng không trả nợ đúng hạn.
CNT_INSTALMENT	Thời hạn của khoản vay tín dụng trước đó của khách hàng, có thể thay đổi theo thời gian. Thời hạn này được tính bằng số tháng và được sử dụng để đánh giá khả năng thanh toán nợ của khách hàng trong quá khứ và hiện tại.	Thông thường, nếu thời hạn của khoản vay trước đó càng dài, thì khách hàng có khả năng thanh toán nợ của mình càng tốt, vì họ đã có kinh nghiệm trong việc quản lý và thanh toán nợ trước đó. Tuy nhiên, nếu thời hạn của khoản vay trước đó quá dài, điều này có thể gây ra áp lực tài chính cho khách hàng, và do đó có thể làm giảm khả năng thanh toán nợ trong tương lai. Thời hạn của khoản vay trước đó cũng có thể thay đổi theo thời gian, ví dụ như trong trường hợp khách hàng chọn gia hạn khoản vay trước đó hoặc thay đổi điều kiện khoản vay. Do đó, thuật ngữ "Term of previous credit (can change over time)" thường được sử dụng để chỉ ra rằng thời hạn của khoản vay trước đó có thể thay đổi

		theo thời gian và không cố định. Đánh giá khả năng thanh toán nợ của khách hàng trong quá khứ và hiện tại, và từ đó giúp đưa ra quyết định về việc cấp vay tín dụng cho khách hàng.
CNT_INSTALMENT_FUTURE	Số lượng khoản trả góp còn lại phải trả trên khoản vay tín dụng trước đó của khách hàng.	Khi một ngân hàng hoặc tổ chức tín dụng cấp vay tiền cho khách hàng, họ thường xem xét lịch sử tín dụng của khách hàng để đánh giá khả năng thanh toán nợ. Nếu khách hàng vẫn còn nhiều khoản trả góp phải trả trên khoản vay tín dụng trước đó, điều này có thể tăng nguy cơ cho ngân hàng hoặc tổ chức tín dụng trong việc cho KH vay tiền. Ngược lại, nếu khách hàng chỉ còn một số lượng nhỏ các khoản trả góp còn lại phải trả trên khoản vay trước đó, điều này có thể cho thấy rằng khách hàng có khả năng thanh toán nợ tốt và có thể được xem xét cho vay tiền.
NAME_CONTRACT_STATUS	Tình trạng hợp đồng của các khoản vay tín dụng của khách hàng trong mỗi tháng.	Các giá trị trong cột này bao gồm: "Active": Hợp đồng đang hoạt động và khách hàng đang tiếp tục thanh toán nợ. "Approved": Hợp đồng đã được phê duyệt và tiền đã được chuyển đến tài khoản của khách hàng. "Canceled": Hợp đồng đã bị hủy bỏ. "Completed": Hợp đồng đã được hoàn thành và

		tất cả các khoản nợ đã được thanh toán. "Demand": Khách hàng đang yêu cầu gia hạn hoặc thay đổi hợp đồng. "Returned To the store": Sản phẩm đã được trả lại cho cửa hàng. "Signed": Hợp đồng đã được ký kết và đang đợi xử lý. Nếu khách hàng có tình trạng hợp đồng "Active" hoặc "Completed" trong tháng đó, điều này cho thấy rằng khách hàng có khả năng thanh toán nợ tốt và có thể được xem xét cho vay tiền. Ngược lại, nếu khách hàng có tình trạng hợp đồng "Canceled" hoặc "Demand", điều này có thể tăng nguy cơ cho ngân hàng hoặc tổ chức tín dụng trong việc cho vay tiền cho khách hàng.
SK_DPD	Số ngày quá hạn trong tháng đối với khoản nợ tín dụng trước đó của khách hàng. Thông thường, DPD được tính bằng cách so sánh ngày thanh toán cuối cùng của khoản nợ trước đó với ngày đáo hạn của khoản nợ đó.	Nếu khách hàng có DPD cao trong tháng đó,điều này cho thấy rằng khách hàng đã trễ hạn thanh toán nợ và có thể làm tăng nguy cơ cho ngân hàng hoặc tổ chức tín dụng trong việc cho vay tiền cho khách hàng. Đánh giá tình trạng tín dụng của khách hàng và xác định mức độ rủi ro trong việc cho vay tiền cho khách hàng. Nếu khách hàng có DPD cao, ngân hàng hoặc tổ chức tín dụng có thể yêu cầu khách hàng thanh toán nợ trước khi xem xét cho vay tiền mới.

		Nếu khoản nợ trước đó có ngày đáo hạn là ngày 10 của tháng và khách hàng đã thanh toán vào ngày 20 của tháng đó, thì DPD sẽ là 10 ngày. Nếu khách hàng không thanh toán đến ngày 30 của tháng đó, thì DPD sẽ là 20 ngày.
SK_DPD_DEF	số ngày trễ thanh toán mặc định (Days Past Due) trong tháng với mức dung sai (tolerance) được áp dụng (các khoản nợ có số tiền vay thấp được bỏ qua) của khoản tín dụng trước đó.	

Bảng 4.4. Bảng khám phá và mô tả các biến trong bảng dữ liệu "POS_CASH_balance"

4.2.2.1. Kiểm tra và xem xét các giá trị rỗng

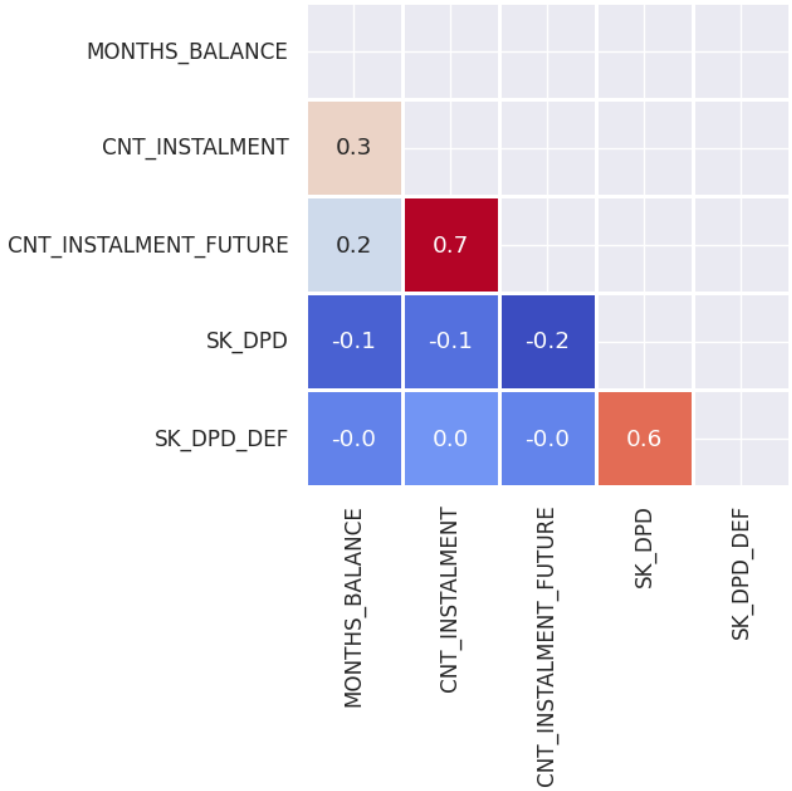
Các thuộc tính trong bảng "POS_CASH_balance" có giá trị rỗng được trình bày trong hình dưới:

Cột "CNT_INSTALMENT_FUTURE" chứa các giá trị rỗng (xấp xỉ 26,08%) khi khách hàng chưa trả các khoản trả góp còn lại trên khoản vay tín dụng trước đó của họ đã được tính toán. Cụ thể, nếu khách hàng không hoàn thành việc trả góp đầy đủ trên khoản vay trước đó, hoặc nếu khách hàng có thể đã hoàn thành toàn bộ kỳ trả nợ dự kiến và không còn kỳ trả nợ còn lại., thì cột "CNT_INSTALMENT_FUTURE" sẽ chứa giá trị missing.

Mối tương quan cao (0.7) giữa cột "CNT_INSTALMENT_FUTURE" và cột "CNT_INSTALMENT" cho thấy sự tương quan mạnh mẽ giữa số lượng kỳ trả nợ còn lại và thời hạn khoản vay tín dụng trước đó. Điều này có ý nghĩa là khi thời hạn khoản vay tín dụng trước đó tăng lên, số lượng kỳ trả nợ còn lại cũng tăng theo và ngược lại.

Cột “CNT_INSTALMENT” chứa các giá trị missing (xấp xỉ 26,07%) khi thông tin về thời hạn khoản vay trước đó của khách hàng không có sẵn trong bảng dữ liệu “POS_CASH_balance”. Điều này có thể xảy ra khi khách hàng không có khoản vay tín dụng trước đó, hoặc khi thông tin về khoản vay trước đó không được cập nhật trong bảng dữ liệu này.

Mối tương quan giữa cột "CNT_INSTALMENT" và "MONTHS_BALANCE" là 0.3 cho thấy một mức tương quan tương đối yếu giữa thời hạn khoản vay tín dụng trước đó và tháng cập nhật thông tin về số dư tài khoản. Điều này có thể ám chỉ rằng sự thay đổi trong thời hạn khoản vay không có mối liên kết mạnh với thời điểm cập nhật thông tin số dư tài khoản.



Hình 4.12. Biểu đồ thể hiện tương quan giữa các biến trong bảng "POS_CASH_balance"

(Nguồn: Nhóm tác giả)

	Total	**Percent**
CNT_INSTALMENT_FUTURE	26087	0.260835

TP.HCM, ngày 15 tháng 6 năm 2023

CNT_INSTALMENT	26071	0.260675

Bảng 4.5. Bảng mô tả tỷ lệ missing data theo từng thuộc tính trong bảng "POS_CASH_balance"

4.2.2.2. Phân tích các biến số

	MONTHS_BALANCE	**CNT_INSTALMENT**	**CNT_INSTALMENT_FUTURE**	**SK_DPD**	**SK_DPD_DEF**
mean	-35.012588	17.089650	10.483840	11.606928	0.654468
std	26.066570	11.995056	11.109058	132.714043	32.762491
min	-96.000000	1.000000	0.000000	0.000000	0.000000
1%	-94.000000	4.000000	0.000000	0.000000	0.000000
10%	-77.000000	6.000000	0.000000	0.000000	0.000000
25%	-54.000000	10.000000	3.000000	0.000000	0.000000
50%	-28.000000	12.000000	7.000000	0.000000	0.000000
75%	-13.000000	24.000000	14.000000	0.000000	0.000000
80%	-11.000000	24.000000	17.000000	0.000000	0.000000
90%	-6.000000	36.000000	24.000000	0.000000	0.000000
95%	-4.000000	45.000000	35.000000	0.000000	0.000000
99%	-2.000000	60.000000	53.000000	235.000000	1.000000
max	-1.000000	92.000000	85.000000	4231.000000	3595.000000

Bảng 4.6. Bảng thống kê mô tả các biến trong bảng dữ liệu "POS_CASH_balance"

Thuộc tính MONTHS_BALANCE:

Thời điểm cập nhật thông tin về số dư tài khoản của khách hàng thường xảy ra trong khoảng từ 35 tháng trước thời điểm đăng ký vay. Điều này cho thấy thông tin về số dư tài khoản được cập nhật tương đối khá gần với thời điểm xem xét hồ sơ.

Giá trị thấp nhất của cột "MONTHS_BALANCE" là -96, cho thấy thông tin về số dư tài khoản có thể được cập nhật lâu nhất là 96 tháng trước thời điểm đăng ký vay. Điều này có thể tạo ra một lỗ hổng thông tin và ảnh hưởng đến độ tin cậy của dữ liệu.

Độ lệch chuẩn (std) của cột "MONTHS_BALANCE" là 26.066570, cho thấy sự biến động lớn về thời điểm cập nhật thông tin số dư tài khoản giữa các khách hàng. Điều này có thể ảnh hưởng đến việc đánh giá rủi ro nợ tài chính và quản lý thu hồi nợ hiệu quả.

Thuộc tính CNT_INSTALMENT

Thời hạn trung bình của khoản vay tín dụng trước đó của khách hàng là 17.089650 tháng. Điều này cho thấy trung bình khách hàng đã có khoảng 17 tháng để thanh toán nợ trên khoản vay trước đó trước khi nộp đơn xin vay mới.

Giá trị thấp nhất của cột "CNT_INSTALMENT" là -1, có thể chỉ ra sự thiếu thông tin hoặc lỗi trong dữ liệu. Việc có giá trị âm không phù hợp trong ngữ cảnh này có thể ảnh hưởng đến độ tin cậy của dữ liệu và đánh giá rủi ro nợ tài chính. Giá trị cao nhất là 92, cho thấy sự đa dạng về thời hạn của khoản vay tín dụng trước đó của khách hàng.

Độ lệch chuẩn (std) của cột "CNT_INSTALMENT" là 11.995056, cho thấy sự biến động trong thời hạn của các khoản vay trước đó giữa các khách hàng. Sự khác biệt này có thể làm tăng rủi ro nợ tài chính của ngân hàng nếu thông tin này không được cập nhật đồng bộ và chính xác đối với từng khách hàng.

Thuộc tính CNT_ INSTALMENT_ FUTURE:

Trung bình số lượng khoản trả góp còn lại phải trả trên khoản vay tín dụng trước đó của khách hàng là 10.483840. Điều này cho thấy trung bình khách hàng còn lại khoảng 10 khoản trả góp để thanh toán trên khoản vay trước đó. Số lượng khoản trả góp còn lại này có thể ảnh hưởng đến khả năng thanh toán nợ và quản lý thu hồi nợ.

Giá trị thấp nhất của cột "CNT_INSTALMENT_FUTURE" là 0, có nghĩa là có khách hàng không còn khoản trả góp nào phải thanh toán trên khoản vay trước đó. Điều này có thể đồng nghĩa với việc khách hàng đã hoàn thành việc thanh toán khoản vay. Giá trị cao nhất là 85, cho thấy sự đa dạng về số lượng khoản trả góp còn lại giữa các khách hàng.

Độ lệch chuẩn là 11.109058, cho thấy sự biến động trong số lượng khoản trả góp còn lại giữa các khoản vay trước đó của khách hàng khác nhau.

Thuộc tính SK_DPD:

Trung bình số ngày quá hạn trong tháng là 11.606928. Điều này cho thấy trung bình khách hàng vượt quá số ngày đáo hạn trên khoản nợ tín dụng trước đó là khoảng 11 ngày trong mỗi tháng. Số ngày quá hạn này có thể là dấu hiệu cho khả năng thanh toán nợ không đủ kịp thời hoặc không đảm bảo đúng hẹn, tạo ra rủi ro nợ tài chính.

Giá trị lớn nhất của số ngày quá hạn trong tháng đối với khoản nợ tín dụng trước đó là 4231 ngày, cho thấy rằng có một số khách hàng gặp khó khăn trong việc thanh toán nợ và có thể ảnh hưởng đến việc thu hồi nợ của ngân hàng. Đồng thời, cho thấy sự đa dạng và biến động mạnh về số ngày quá hạn trong tháng giữa các khoản nợ tín dụng trước đó của khách hàng. Sự biến động này có thể tăng rủi ro nợ tài chính và tạo ra khó khăn trong việc thu hồi nợ hiệu quả.

Độ lệch chuẩn (std) của cột "SK_DPD" là 132.714043, cho thấy mức độ biến động lớn về số ngày quá hạn trong tháng. Điều này yêu cầu một quy trình quản lý rủi ro nợ tài chính chặt chẽ và chiến lược thu hồi nợ linh hoạt để đảm bảo khả năng thanh toán nợ và giảm thiểu tác động tiêu cực của việc quá hạn.

Thuộc tính SK_DPD_DEF:

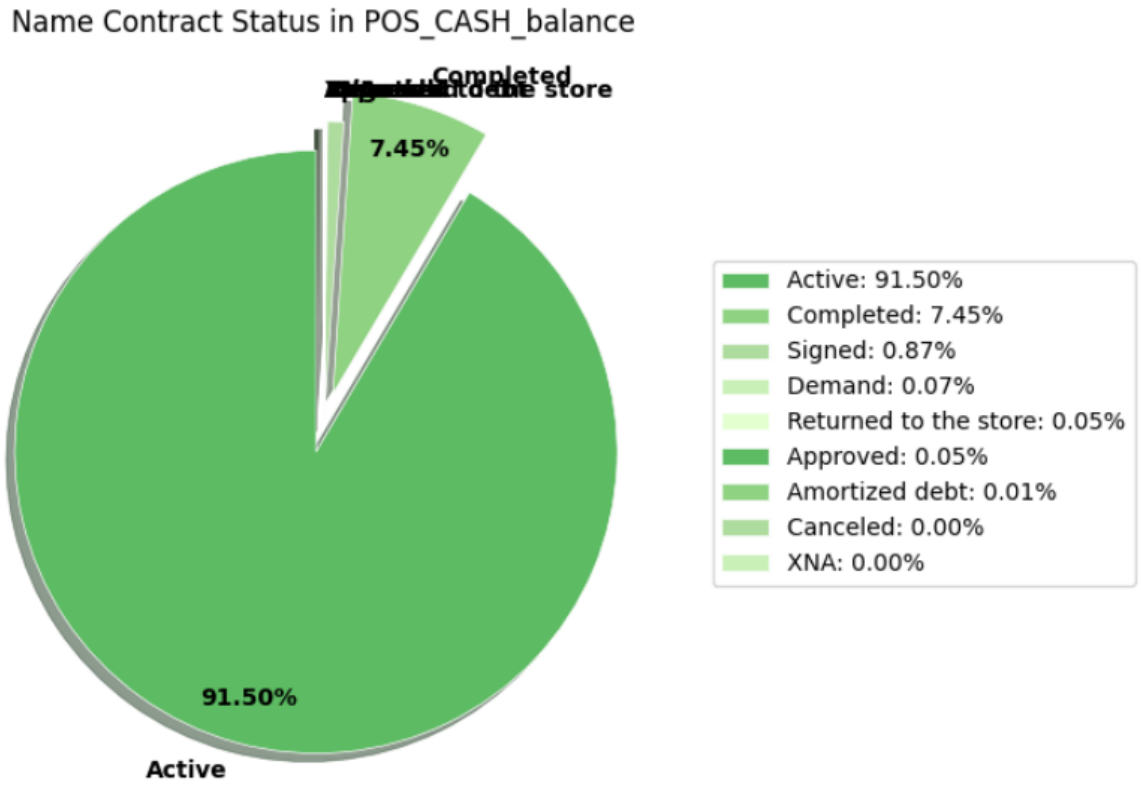
Giá trị trung bình (mean) của cột “SK_DPD_DEF” 0.654468 cho thấy, trung bình mỗi khoản nợ tín dụng trước đó của khách hàng ghi nhận khoảng 0.65 ngày quá hạn tích lũy trong tháng, số này cho thấy rằng số ngày quá hạn trung bình của khoản nợ tín dụng không quá cao.

TP.HCM, ngày 15 tháng 6 năm 2023

Giá trị nhỏ nhất (min) là 0 cho thấy một số khoản nợ không ghi nhận số ngày quá hạn tích lũy trong tháng. Điều này có thể chỉ ra tình huống tốt hơn trong việc quản lý nợ và thanh toán đúng hẹn. Tuy nhiên, giá trị lớn nhất của cột SK_DPD_DEF là 3595 ngày, cho thấy rằng có một số khách hàng đã quá hạn thanh toán trong một khoảng thời gian dài và có thể ảnh hưởng đến việc thu hồi nợ của ngân hàng.

Độ lệch chuẩn (std) là 32.762491, khá cao cho thấy sự khác biệt lớn giữa số ngày quá hạn của các khoản nợ tín dụng khác nhau. Việc này có thể ảnh hưởng đến khả năng đánh giá rủi ro nợ tài chính và thu hồi nợ hiệu quả của ngân hàng.

4.2.2.3. Phân tích biến phân loại NAME_CONTRACT_STATUS



Hình 4.13. Biểu đồ thể hiện phần trăm các trạng thái của hợp đồng

(Nguồn: Nhóm tác giả)

Biểu đồ phần trăm các trạng thái của hợp đồng cho thấy hầu hết các hợp đồng trong bảng POS_CASH_balance đang ở trạng thái active, chiếm tới 91.50%. Sau đó là tỷ lệ các hợp đồng đã được hoàn thành hoặc đang ở các trạng thái khác.

Thông qua phân tích biểu đồ, ta có thể phân nhóm khách hàng dựa trên trạng thái hợp đồng. Các khách hàng có hợp đồng ở trạng thái active có thể được coi là nhóm ổn định, trong khi các khách hàng có hợp đồng ở các trạng thái khác có thể được coi là nhóm có rủi ro cao hơn.

Để giảm thiểu rủi ro cho tổ chức cho vay, đề xuất điều chỉnh lãi suất phù hợp có thể được áp dụng cho các khách hàng thuộc nhóm có rủi ro cao hơn.

4.2.3. Kết luận

MONTHS_BALANCE càng lớn lớn chứng tỏ khách hàng đã sử dụng các tài khoản trong một khoảng thời gian dài. Do đó khách hàng có khả năng quản lý tài chính tốt hơn đồng thời có thể có khả năng trả nợ đúng hạn.

Ngược lại, MONTHS_BALANCE nhỏ cho thấy khách hàng đã sử dụng các tài khoản trong một khoảng thời gian ngắn. Từ đó thể hiện khách hàng có khả năng không quản lý tài chính tốt đồng thời có thể có khả năng không trả nợ đúng hạn.

CNT_INSTALMENT càng dài cho thấy khách hàng có khả năng thanh toán nợ của mình càng tốt, vì họ đã có kinh nghiệm trong việc quản lý và thanh toán nợ trước đó.

Tuy nhiên, điều này có thể gây ra áp lực tài chính cho khách hàng và có thể làm giảm khả năng thanh toán nợ trong tương lai. Thuộc tính này đánh giá khả năng thanh toán nợ của khách hàng trong quá khứ và hiện tại và giúp tổ chức đưa ra quyết định về việc cấp vay tín dụng cho khách hàng.

CNT_INSTALMENT_FUTURE càng lớn thể hiện khách hàng vẫn còn nhiều khoản trả góp phải trả trên khoản vay tín dụng trước đó. Từ đó, tăng nguy cơ cho ngân hàng hoặc tổ chức tín dụng trong việc cho KH vay tiền.

Ngược lại, CNT_INSTALMENT_FUTURE càng nhỏ cho biết khách hàng có khả năng thanh toán nợ tốt và có thể được xem xét cho vay tiền.

NAME_CONTRACT_STATUS có giá trị "Active" hoặc "Completed" thể hiện khách hàng có khả năng thanh toán nợ tốt và có thể được xem xét cho vay tiền.

Ngược lại, nếu NAME_CONTRACT_STATUS có giá trị "Canceled" hoặc "Demand" cho thấy có thể có nguy cơ cho ngân hàng hoặc tổ chức tín dụng trong việc cho vay tiền cho khách hàng.

4.3. Bảng previous_application

4.3.1. Tổng quan về bảng

Bảng "previous_application" trong Home Credit là một bảng dữ liệu chứa thông tin về các đơn đăng ký vay trước đó của khách hàng. Bảng bao gồm 1.670.214 dòng, 37 cột, tương ứng với 1670214 quan sát là các giá trị duy nhất, thể hiện các khoản vay trước mà liên quan tới khoản vay hiện tại và 37 thuộc tính cho biết các thông tin về những khoản vay trước đó.

Bảng chứa 21 biến là dữ liệu số, cung cấp các thông tin của các khoản vay trước, về khoản trả góp hàng tháng của khách hàng, số tiền khách hàng yêu cầu vay trong đơn xin, lãi suất cơ bản, khoảng thời gian tính từ ngày nộp đơn vay hiện tại đến ngày đầu tiên mà khoản vay trước đó đáo hạn,...và 16 biến là dữ liệu phân loại, ghi nhận các thông tin như các loại hợp đồng sản phẩm, các loại danh mục sản phẩm, các lý do từ chối đơn đăng ký vay,...

4.3.2. Khám phá các biến

Thuộc tính	**Mô tả**	**Các thuộc tính**
SK_ID_PREV	ID của khoản vay trước liên quan đến khoản vay hiện tại (Một khoản vay hiện tại có thể có 0, 1, 2 hoặc nhiều khoản vay trước ở Home Credit. Do đó, SK_ID_CURR)	
SK_ID_CURR	ID của khoản vay trong mẫu	

NAME_CONTRACT_TYPE	Loại sản phẩm hợp đồng (Cash loans, consumer loans, Revolving loans) của khoản vay trước	- Cash loans (Khoản vay tiền mặt): Đây là loại hợp đồng vay thông thường trong đó khách hàng được cấp một số tiền cụ thể và phải trả lại theo lịch trả góp được định sẵn. - Consumer loans (Khoản vay tiêu dùng): Đây là loại hợp đồng vay được sử dụng để tài trợ các nhu cầu tiêu dùng của khách hàng, chẳng hạn như mua sắm, du lịch, nội thất nhà cửa, hoặc chi tiêu hàng ngày,… - Revolving loans (Credit card): Đại diện cho các khoản vay có tính tái sử dụng. Trong loại hợp đồng này, khách hàng có một khoản tín dụng tối đa đã được phê duyệt và có thể mượn và trả nợ lặp lại trong phạm vi tín dụng đã được xác định. Số tiền trả hàng tháng được tính dựa trên số dư chưa trả và tỷ lệ phần trăm được xác định trước. - "XNA": Đây là giá trị không rõ ràng hoặc không xác định (Unknown). Thông thường, giá trị "XNA" được sử dụng khi dữ liệu bị thiếu hoặc không thể xác định chính xác loại hợp đồng vay. Đây có thể là một giá trị bất thường hoặc các trường hợp đặc biệt trong dữ liệu.
AMT_ANNUITY	Khoản trả góp hàng tháng của khoản vay trước	- Giá trị thể hiện khoản trả góp hàng tháng của khoản vay trước rỗng khi đơn xin vay này bị hủy hoặc từ chối hoặc trong hầu hết các đơn xin vay có trạng thái “Unused offer”.

TP.HCM, ngày 15 tháng 6 năm 2023

AMT_APPLICATION	Số tiền tín dụng mà khách hàng yêu cầu trong đơn xin vay trước	Số tiền tín dụng mà khách hàng yêu cầu trong đơn xin vay trước có giá trị bằng 0 khi đơn xin vay này bị hủy hoặc từ chối.
AMT_CREDIT	Số tiền tín dụng cuối cùng được chấp nhận trong đơn xin trước. Số này có thể khác so với số tiền thể hiện ở cột AMT_APPLICATION vì "AMT_APPLICATION" là số tiền mà khách hàng đầu tiên yêu cầu vay, nhưng trong quá trình xét duyệt, khách hàng có thể nhận được một số tiền khác, là "AMT_CREDIT".	- Số tiền được cấp (đã bao gồm bảo hiểm nếu trước đó khách hàng cũng yêu cầu cung cấp cả khoản vay cho phí bảo hiểm (Được thể hiện với giá trị bằng 1 ở cột NFLAG_INSURED_ON_APPROVAL). Số tiền tín này được tính toán theo công thức sau: AMT_CREDIT = AMT_GOODS_PRICE + insurance. - Tương tự "AMT_APPLICATION", số tiền tín dụng cuối cùng được chấp nhận trong đơn xin trước có giá trị bằng 0 khi đơn xin vay này bị hủy hoặc từ chối.
AMT_DOWN_PAYMENT	Số tiền trả trước cho đơn trước đó	
AMT_GOODS_PRICE	Giá hàng hóa mà khách hàng yêu cầu vay (nếu có) trong đơn vay trước đó.	- Giá trị thể hiện giá hàng hóa mà khách hàng yêu cầu vay (nếu có) trong đơn vay trước đó rỗng khi đơn xin vay này bị hủy hoặc từ chối. - Giá hàng hóa mà khách hàng yêu cầu vay bằng giá trị số tiền tín dụng mà khách hàng yêu cầu trong đơn xin vay trước.
WEEKDAY_APPR_PROCESS_START	Ngày trong tuần mà khách hàng nộp đơn xin trước đó.	
HOUR_APPR_PROC	Giờ xấp xỉ trong ngày mà	

ESS_START	khách hàng nộp đơn xin trước đó.	
FLAG_LAST_APPL_PER_CONTRACT	Đánh dấu nếu đây là đơn xin cuối cùng cho hợp đồng trước đó. Trong một số trường hợp, có thể có nhiều đơn xin cho một hợp đồng duy nhất (bởi nhầm lẫn của nhân viên).	- Giá trị Y: Đây là đơn xin cuối cùng cho hợp đồng trước đó. - Giá trị N: Đây không là đơn xin cuối cùng cho hợp đồng trước đó. Trong bộ dữ liệu Home Credit, mỗi hợp đồng vay có thể liên quan đến nhiều đơn xin vay. Giá trị "N" được sử dụng để đánh dấu các đơn xin vay trước đó trong hợp đồng, trong khi giá trị "Y" sẽ chỉ định đây là đơn xin vay cuối cùng trong hợp đồng. Điều này giúp phân biệt các đơn xin vay trong cùng một hợp đồng và có thể hữu ích trong việc phân tích dữ liệu và đối chiếu thông tin giữa các đơn xin vay khác nhau.
NFLAG_LAST_APPL_IN_DAY	Đánh dấu nếu đây là đơn xin cuối cùng trong ngày của khách hàng. Đôi khi, khách hàng nộp nhiều đơn xin trong cùng một ngày và hiếm khi, có thể tồn tại các đơn xin trùng lặp trong hệ thống.	- Giá trị 0: Điều này cho thấy đơn xin vay không phải là đơn cuối cùng được xử lý trong ngày đó. Có nghĩa là có ít nhất một đơn xin vay khác đã được xử lý sau đó trong cùng một ngày. - Giá trị 1: Điều này cho thấy đơn xin vay là đơn cuối cùng được xử lý trong ngày đó. Không có đơn xin vay khác được xử lý sau đó trong cùng một ngày. - Giá trị này có thể hữu ích trong việc theo dõi và phân tích thông tin về các đơn xin vay trong cùng một ngày và xác định xem liệu đơn xin vay đó có phải là cuối cùng được xử lý trong

		ngày hay không.
NFLAG_MICRO_CASH	Đánh dấu nếu là một khoản vay tài chính nhỏ. (Micro finance loan)	- Giá trị 0: Điều này cho thấy đơn xin vay không liên quan đến khoản vay nhỏ. Có nghĩa là khoản vay trong đơn xin vay không thuộc vào danh mục khoản vay nhỏ. - Giá trị 1: Điều này cho thấy đơn xin vay liên quan đến khoản vay nhỏ. Có nghĩa là khoản vay trong đơn xin vay thuộc vào danh mục khoản vay nhỏ, thường là các khoản vay có số tiền nhỏ hơn và thời hạn ngắn hơn so với các khoản vay thông thường. - Giá trị này có thể hữu ích để xác định xem đơn xin vay có liên quan đến khoản vay nhỏ hay không, và có thể sử dụng để phân loại và phân tích các loại khoản vay khác nhau trong bộ dữ liệu.
RATE_DOWN_PAYMENT	Tỷ lệ trả trước được chuẩn hóa trên tín dụng trước đó.	- Giá trị thể hiện tỷ lệ trả trước được chuẩn hóa trên tín dụng trước đó rỗng khi giá trị thể hiện số tiền trả trước cho đơn trước đó tương ứng rỗng. - Giá trị này rỗng khi giá trị thể hiện số tiền trả trước cho đơn trước đó tương ứng rỗng.
RATE_INTEREST_PRIMARY	Lãi suất cơ bản được chuẩn hóa trên tín dụng trước đó. (Lãi suất chuẩn hoặc lãi suất cố định mà khách hàng phải	

	trả cho khoản vay hoặc hợp đồng tín dụng.)	
RATE_INTEREST_PRIVILEGED	Lãi suất ưu đãi được chuẩn hóa trên tín dụng trước đó (lãi suất ưu đãi). (Lãi suất đặc biệt hoặc ưu đãi được áp dụng cho một nhóm khách hàng đặc biệt hoặc có đặc quyền.)	
NAME_CASH_LOAN_PURPOSE	Mục đích của khoản vay tiền mặt trong đơn xin trước đó.	
NAME_CONTRACT_STATUS	Trạng thái hợp đồng của đơn xin trước đó (được chấp thuận, đã hủy, ...).	- Approved: Đây là trạng thái của hợp đồng khi nó đã được duyệt và tiến hành. Trong trường hợp này, người vay đã nhận được khoản vay và quá trình vay đã được thực hiện thành công. - Canceled: Trạng thái "Canceled" đại diện cho việc hủy bỏ hợp đồng trước khi nó được thực hiện. Người vay hoặc bên cho vay đã quyết định không tiến hành vay và hợp đồng đã bị hủy. - Refused: Khi một hợp đồng được từ chối, trạng thái "Refused" được sử dụng. Điều này có nghĩa là người vay đã nộp đơn và yêu cầu vay, nhưng họ không được chấp thuận và không nhận được khoản vay.

		- Unused offer: "Unused offer" ám chỉ rằng một đề nghị vay tiền đã được cung cấp cho người vay, nhưng người vay không sử dụng hoặc không chấp nhận đề nghị đó. Điều này có thể xảy ra khi người vay không quan tâm đến việc vay tiền hoặc tìm kiếm các lựa chọn khác.
DAYS_DECISION	Số ngày so với đơn xin hiện tại khi quyết định về đơn xin trước đó được đưa ra.	- Giá trị dương: Đây là số ngày sau ngày đệ trình đơn xin vay cho đến ngày quyết định. Ví dụ: 30, 60, 90, ... Giá trị dương thường đại diện cho việc công ty mất một khoảng thời gian xử lý đơn xin vay trước khi đưa ra quyết định. - Giá trị 0: Điều này cho thấy quyết định về đơn xin vay đã được đưa ra trong ngày đệ trình đơn xin vay. - Giá trị âm: Đây là số ngày trước ngày đệ trình đơn xin vay mà quyết định được đưa ra. Ví dụ: -30, -60, -90, ... Giá trị âm thường đại diện cho việc công ty đã đưa ra quyết định trước ngày đệ trình đơn xin vay.
NAME_PAYMENT_TYPE	Phương thức thanh toán mà khách hàng chọn để thanh toán cho đơn xin trước đó.	- Cash through the bank: Đây là phương thức thanh toán bằng tiền mặt thông qua ngân hàng. Người vay trả tiền mặt tại quầy giao dịch của ngân hàng hoặc thông qua dịch vụ chuyển khoản tiền mặt của ngân hàng.

		- XNA: Giá trị "XNA" thường được sử dụng khi không có thông tin cụ thể về phương thức thanh toán hoặc không xác định được phương thức thanh toán được sử dụng. Đây có thể là các trường hợp đặc biệt hoặc dữ liệu thiếu. - Non-cash from your account: Đây là phương thức thanh toán không sử dụng tiền mặt, mà thông qua việc trừ tiền từ tài khoản ngân hàng của người vay. Người vay cho phép ngân hàng trừ tiền từ tài khoản của họ để thanh toán khoản vay. - Cashless from the account of the employer: Đây là phương thức thanh toán không sử dụng tiền mặt, mà thông qua việc người cho vay trừ tiền từ tài khoản ngân hàng của người vay. Người cho vay trừ tiền từ tài khoản của người vay để thanh toán khoản vay.
CODE_REJECT_REASON	Lý do tại sao đơn xin trước đó bị từ chối.	- XAP: Đơn xin vay bị từ chối vì không thuộc danh sách các đơn xin vay hợp lệ hoặc không được phê duyệt vì một số lý do như: người vay chưa đủ tuổi, không có tài sản đảm bảo, thu nhập không đủ, lịch sử tín dụng xấu,... - HC: Đơn xin vay bị từ chối vì liên quan đến việc tìm hiểu thông tin và kiểm tra tín dụng của khách hàng. Kiểm tra tín dụng thường bao

		gồm cả việc kiểm tra điểm tín dụng và thông tin khác liên quan đến khách hàng. Điểm tín dụng là một chỉ số được sử dụng để đánh giá khả năng và tính đáng tin cậy trong việc trả nợ của một cá nhân hoặc một tổ chức. Điểm tín dụng thường được tính toán dựa trên thông tin tài chính và lịch sử tín dụng của người vay, bao gồm việc thanh toán các khoản vay trước đó, các khoản nợ hiện tại, lịch sử tín dụng, số lần truy cập tín dụng, và các yếu tố khác. Các tổ chức tín dụng, như ngân hàng hoặc công ty tài chính, sử dụng điểm tín dụng để đưa ra quyết định về việc xét duyệt đơn xin vay hoặc cung cấp dịch vụ tài chính. Điểm tín dụng cao thường được coi là một chỉ số tốt, cho thấy khả năng trả nợ đúng hẹn và đáng tin cậy của người vay. Ngược lại, điểm tín dụng thấp hoặc không đạt yêu cầu có thể là một trong những yếu tố bất lợi trong việc xét duyệt đơn vay hoặc dẫn đến từ chối đơn xin vay. Mỗi tổ chức có thể có các phương pháp tính điểm tín dụng khác nhau, và các yếu tố quyết định cụ thể cũng có thể khác nhau. Tuy nhiên, điểm tín dụng thường được biểu thị bằng một số hoặc hạng mức nhất định, thường từ 300 đến 850 điểm. Bên cạnh đó, kiểm tra tín dụng cũng liên quan đến việc xác minh thông tin, tài liệu hoặc yếu tố khác liên quan đến khả năng thanh toán và khả năng trả nợ của khách hàng. Điều này

		có thể bao gồm việc kiểm tra thông tin về thu nhập, lịch sử tài chính, khả năng trả nợ, hoặc các yếu tố khác mà tổ chức tín dụng hoặc công ty tài chính quan tâm để đảm bảo rằng khoản vay được cung cấp là an toàn và có khả năng trả lại trong tương lai. - LIMIT: Đơn xin vay bị từ chối vì vượt quá giới hạn tín dụng hoặc giới hạn tín dụng không đủ để đáp ứng yêu cầu vay. Giới hạn tín dụng là một khoản tiền tối đa mà người vay được cho phép vay từ tổ chức tín dụng hoặc công ty tài chính. Đây là một hạn mức tín dụng được đặt ra dựa trên nhiều yếu tố, bao gồm thu nhập, lịch sử tín dụng, khả năng trả nợ và các yếu tố tài chính khác của người vay. Khi một đơn xin vay vượt quá giới hạn tín dụng đã được đặt ra hoặc giới hạn tín dụng không đủ để đáp ứng yêu cầu vay, đơn xin vay có thể bị từ chối. Điều này có thể xảy ra khi số tiền vay yêu cầu vượt quá khả năng trả nợ của người vay hoặc khi người vay đã sử dụng hết giới hạn tín dụng hiện có. Khi khách hàng sử dụng hết giới hạn tín dụng, có thể có một số lý do như: Khách hàng đã sử dụng tất cả số tiền cho vay có sẵn trong tài khoản hoặc thẻ tín dụng của họ; Khách hàng đang có các khoản vay đang còn nợ, và việc này giảm đi số tiền khả dụng trong giới hạn tín dụng…

TP.HCM, ngày 15 tháng 6 năm 2023

		- SCO: Đơn xin vay bị từ chối do điểm tín dụng không đạt yêu cầu hoặc không đủ để đáp ứng tiêu chuẩn vay của công ty. - SCOFR: Đơn xin vay bị từ chối do điểm tín dụng không đạt yêu cầu hoặc không đủ để đáp ứng tiêu chuẩn vay của công ty dựa trên việc tìm hiểu thông tin. Điều này cho thấy không chỉ dựa trên điểm tín dụng, mà công ty cũng xem xét các thông tin khác để đánh giá khả năng trả nợ của người vay. Dựa trên việc tìm hiểu thông tin, công ty đã kết luận rằng người vay không đáp ứng đủ tiêu chuẩn vay dựa trên các yếu tố khác ngoài điểm tín dụng. - SYSTEM: Đơn xin vay bị từ chối bởi hệ thống tự động, có thể do lỗi kỹ thuật hoặc các vấn đề liên quan đến xử lý đơn hàng tự động. - VERIF: Đơn xin vay bị từ chối do không thể xác minh thông tin khách hàng, ví dụ như thông tin cá nhân không chính xác hoặc không thể xác minh. - CLIENT: Đơn xin vay bị từ chối do các vấn đề liên quan đến khách hàng, bao gồm thông tin không đúng hoặc không đáng tin cậy.

TP.HCM, ngày 15 tháng 6 năm 2023

		- XNA: Thường được sử dụng để đại diện cho các lý do từ chối không xác định, không rõ ràng. không đầy đủ. Điều này có thể xảy ra khi không có thông tin cụ thể hoặc không có lý do cụ thể được ghi lại cho việc từ chối đơn xin vay.
NAME_TYPE_SUITE	Người đi kèm khi khách hàng nộp đơn xin vay	- "Unaccompanied": Khách hàng nộp đơn một mình, không có ai đi kèm. - "Spouse, partner": Người đi kèm là vợ/chồng hoặc đối tác của khách hàng. - "Family": Người đi kèm là thành viên trong gia đình của khách hàng. - "Children": Người đi kèm là con cái của khách hàng. - "Other_A": Người đi kèm khác (không rõ ràng trong dữ liệu). - "Group of people": Nhóm người đi kèm, có thể là một nhóm người hoặc đại diện một tổ chức. - "Other_B": Người đi kèm khác (không rõ ràng trong dữ liệu, loại khác so với "Other_A").
NAME_CLIENT_TYPE	Khách hàng là khách hàng cũ hay mới khi nộp đơn xin trước đó.	- "New": Đại diện cho khách hàng mới, tức là khách hàng chưa từng có hồ sơ vay với Home Credit trước đó. - "Repeater": Đại diện cho khách hàng tái vay, tức là khách hàng đã từng có hồ sơ vay với Home Credit và đang nộp đơn xin vay lần thứ

		hai trở lên. - "Refreshed": Đại diện cho khách hàng cập nhật thông tin, tức là khách hàng đã từng có hồ sơ vay với Home Credit và đang cung cấp lại thông tin mới trong đơn xin vay. - "XNA": Giá trị này được sử dụng khi thông tin về loại khách hàng không có sẵn hoặc không xác định trong dữ liệu.
NAME_GOODS_CATEGORY	Loại hàng hóa mà khách hàng đăng ký trong đơn xin trước đó.	
NAME_PORTFOLIO	Đơn xin trước đó thuộc danh mục nào (CASH, POS, CAR, ...).	- Cột "NAME_PORTFOLIO" giúp hiểu rõ hơn về tính chất cũng như là mục đích sử dụng của các loại sản phẩm hợp đồng vay "NAME_CONTRACT_TYPE".
NAME_PRODUCT_TYPE	Đơn xin trước đó là "x-sell" hay "walk-in".	- X-sell: Đại diện cho loại sản phẩm tài chính được cung cấp cho khách hàng hiện tại của công ty, người dùng đã có quan hệ giao dịch trước đó với công ty. Thông thường, khi khách hàng đã vay một khoản vay trước đó và công ty đề xuất một sản phẩm khác để bán cho khách hàng này, giá trị "x-sell" được sử dụng để chỉ loại sản phẩm này. - "Walk-in": Đại diện cho khách hàng đến trực tiếp địa điểm của công ty để nộp đơn xin vay hoặc tham gia giao dịch mà không thông qua

		các kênh truyền thông hoặc đại lý. Khi khách hàng đến công ty "tự đi bộ" để thực hiện các giao dịch tài chính, giá trị "walk-in" được sử dụng để chỉ loại hình giao dịch này. - "XNA": Giá trị này được sử dụng khi thông tin về loại sản phẩm không có sẵn hoặc không xác định trong dữ liệu.
CHANNEL_TYPE	Kênh thông qua đó tiếp cận được khách hàng trong đơn xin trước đó.	- "Credit and cash offices" (Văn phòng tín dụng và tiền mặt): Đại diện cho việc khách hàng nộp đơn xin vay hoặc giao dịch tại văn phòng của công ty tín dụng. Đây có thể là một kênh trực tiếp để khách hàng làm các thủ tục vay tiền hoặc thanh toán trực tiếp. - "Country-wide" (Toàn quốc): Đại diện cho việc khách hàng gửi đơn xin vay hoặc giao dịch thông qua các kênh phân phối toàn quốc, ví dụ như đại lý, điểm giao dịch hoặc đại lý trên toàn quốc. - "Stone": Đại diện cho việc khách hàng gửi đơn xin vay hoặc giao dịch thông qua kênh Stone - một kênh bán hàng trực tuyến được cung cấp bởi Home Credit. Đây có thể là một ứng dụng di động hoặc trang web cho phép khách hàng nộp đơn xin vay và thực hiện các giao dịch tài chính.

		- "Regional / Local": Đại diện cho việc khách hàng gửi đơn xin vay hoặc giao dịch thông qua các kênh phân phối cấp địa phương hoặc vùng miền. - "Contact center": Đại diện cho việc khách hàng gửi đơn xin vay hoặc giao dịch thông qua trung tâm liên hệ của công ty. Khách hàng có thể liên hệ qua điện thoại, email hoặc các hình thức liên lạc khác để nộp đơn hoặc yêu cầu hỗ trợ về tài chính. - "AP+ (Cash loan)": Đại diện cho việc khách hàng gửi đơn xin vay hoặc giao dịch thông qua kênh AP+ - một dịch vụ cho phép khách hàng nộp đơn và thực hiện các giao dịch tài chính thông qua các điểm chấp nhận tiền mặt. - "Channel of corporate sales": Đại diện cho việc khách hàng gửi đơn xin vay hoặc giao dịch thông qua các kênh bán hàng dành riêng cho khách hàng doanh nghiệp. - "Car dealer": Đại diện cho việc khách hàng gửi đơn xin vay hoặc giao dịch thông qua các đại lý ô tô.
SELLERPLACE_AREA	Selling area of seller place of the previous application	- Giá trị dương: Đại diện cho diện tích khu vực bán hàng được ghi lại trong đơn xin vay trước

		đó. Giá trị này có thể được xác định trong đơn xin vay hoặc được cung cấp bởi người bán hàng hoặc tổ chức liên quan. - Giá trị 0: Đại diện cho trường hợp không có diện tích khu vực bán hàng được ghi lại. Điều này có thể xảy ra khi thông tin về diện tích không có sẵn hoặc không được cung cấp. - Giá trị -1: Đại diện cho trường hợp thông tin về diện tích khu vực bán hàng không có sẵn hoặc không được cung cấp. Giá trị -1 thường được sử dụng để chỉ ra sự thiếu thông tin hoặc không khả thi để thu thập thông tin về diện tích khu vực bán hàng.
NAME_SELLER_INDUSTRY	Ngành nghề của người bán trong đơn xin trước đó. (The industry of the seller)	
CNT_PAYMENT	Kỳ hạn (số lần thanh toán) của tín dụng trước đó tại thời điểm đơn xin trước đó.	
NAME_YIELD_GROUP	Nhóm tỷ lệ lãi suất được phân loại thành nhóm lãi suất nhỏ, trung bình và cao trong đơn xin trước đó.	- Low Action: Nhóm lợi suất Low Action thường áp dụng cho các khoản vay có lợi suất thấp và không yêu cầu nhiều hành động từ khách hàng. - Middle: Nhóm lợi suất Middle thường áp

		dụng cho các khoản vay có mức lợi suất trung bình. - High: Nhóm lợi suất High thường áp dụng cho các khoản vay có mức lợi suất cao hơn so với các nhóm khác. - Low Normal: Nhóm lợi suất Low Normal thường áp dụng cho các khoản vay có lợi suất thấp và có yêu cầu hành động từ khách hàng như việc thực hiện các biện pháp đảm bảo thanh toán đúng hạn.
PRODUCT_COMBINATION	Chi tiết của sản phẩm kết hợp trong đơn xin trước đó. (Detailed product combination of the previous application)	- "Card Street": Đơn xin vay liên quan đến việc sử dụng thẻ tín dụng và giao dịch trực tiếp tại cửa hàng. - "Card X-Sell": Đơn xin vay liên quan đến việc sử dụng thẻ tín dụng và có liên kết với các sản phẩm hoặc dịch vụ khác. - "Cash": Đơn xin vay liên quan đến việc nhận tiền mặt. - "Cash Street: high": Đơn xin vay liên quan đến việc nhận tiền mặt trực tiếp tại cửa hàng, với mức vay cao. - "Cash Street: middle": Đơn xin vay liên quan đến việc nhận tiền mặt trực tiếp tại cửa hàng, với mức vay trung bình. - "Cash Street: low": Đơn xin vay liên quan đến việc nhận tiền mặt trực tiếp tại cửa hàng, với mức vay thấp.

TP.HCM, ngày 15 tháng 6 năm 2023

		- "Cash X-Sell: high": Đơn xin vay liên quan đến việc nhận tiền mặt và có liên kết với các sản phẩm hoặc dịch vụ khác, với mức vay cao. - "Cash X-Sell: middle": Đơn xin vay liên quan đến việc nhận tiền mặt và có liên kết với các sản phẩm hoặc dịch vụ khác, với mức vay trung bình. - "Cash X-Sell: low": Đơn xin vay liên quan đến việc nhận tiền mặt và có liên kết với các sản phẩm hoặc dịch vụ khác, với mức vay thấp. - "POS household without interest": Đơn xin vay liên quan đến việc mua sắm tại cửa hàng tiện lợi hoặc siêu thị, không có lãi suất, trong lĩnh vực gia đình. - "POS mobile with interest": Đơn xin vay liên quan đến việc mua sắm thiết bị di động tại cửa hàng tiện lợi hoặc siêu thị, có lãi suất. - "POS mobile without interest": Đơn xin vay liên quan đến việc mua sắm thiết bị di động tại cửa hàng tiện lợi hoặc siêu thị, không có lãi suất. - "POS industry with interest": Đơn xin vay liên quan đến việc mua sắm trong ngành công nghiệp tại cửa hàng tiện lợi hoặc siêu thị, có lãi suất. - "POS industry without interest": Đơn xin vay liên quan đến việc mua sắm trong ngành công nghiệp tại cửa hàng tiện lợi hoặc siêu thị,

		không có lãi suất. - "POS other with interest": Đơn xin vay liên quan đến việc mua sắm các mặt hàng khác tại cửa hàng tiện lợi hoặc siêu thị, có lãi suất. - "POS other without interest": Đơn xin vay liên quan đến việc mua sắm các mặt hàng khác tại cửa hàng tiện lợi hoặc siêu thị, không có lãi suất.
DAYS_FIRST_DRAWING	Khoảng thời gian tính từ ngày nộp đơn vay hiện tại đến ngày đầu tiên mà khoản vay trước đó được giải ngân.	- Giải ngân là việc ngân hàng hay các tổ chức tài chính thực hiện chi tiền cho bên đi vay để cung ứng vốn cho bên vay sử dụng vào đúng mục đích vay vốn. - Giá trị âm nếu thời gian ở quá khứ. - Các giá trị = 365243 (xấp xỉ 1000 năm) có thể hiểu là thời gian này chưa xác định hoặc là khoản vay trước đó chưa đến thời điểm thanh toán đầu tiên.
DAYS_FIRST_DUE	Khoảng thời gian tính từ ngày nộp đơn vay hiện tại đến ngày đầu tiên mà khoản vay trước đó được yêu cầu thanh toán.	- Giá trị âm nếu thời gian ở quá khứ. - Các giá trị = 365243 (xấp xỉ 1000 năm) có thể hiểu là thời gian này chưa xác định hoặc là khoản vay trước đó chưa đến thời điểm thanh toán đầu tiên.
DAYS_LAST_DUE_1ST_VERSION	Khoảng thời gian tính từ ngày nộp đơn vay hiện tại đến ngày cuối cùng mà khoản vay trước đó cần phải thanh toán (last due) trong phiên bản đầu tiên.	- Vì biến này là một dự đoán, nó có thể có giá trị dương (sự kiện trong tương lai), thường xảy ra khi khách hàng hoàn thành khoản vay trước thời hạn đặt ra ban đầu. - Trong trường hợp của các khoản vay dạng Revolving, giá trị 365243 thường chỉ xuất

		hiện khi hạn mức tín dụng vẫn còn hoạt động. Vì các khoản vay Revolving không có ngày cuối cùng cụ thể để thanh toán, giá trị 365243 được sử dụng để biểu thị rằng khoản vay vẫn đang trong trạng thái hoạt động và không có ngày cuối cùng để thanh toán.
DAYS_LAST_DUE	Khoảng thời gian tính từ ngày nộp đơn vay hiện tại đến ngày cuối cùng mà khoản vay trước đó cần phải thanh toán	- Giá trị âm: Các giá trị âm trong cột này cho biết khoản vay trước đó cần phải được thanh toán trước ngày nộp đơn vay hiện tại. Ví dụ, nếu giá trị là -30, nghĩa là khoản vay trước đó cần phải được thanh toán 30 ngày trước ngày nộp đơn vay hiện tại. - Nếu khoản thanh toán cuối cùng chưa được thanh toán, giá trị sẽ là 365243.
DAYS_TERMINATION	Số ngày so với ngày nộp đơn xin hiện tại khi ngày dự kiến kết thúc của đơn xin trước đó.	- Giá trị âm nếu thời gian ở quá khứ. - Các giá trị = 365243 (xấp xỉ 1000 năm) có thể hiểu là thời gian này chưa xác định hoặc là khoản nợ vẫn hoạt động ở hiện tại.
NFLAG_INSURED_ON_APPROVAL	Khách hàng có yêu cầu bảo hiểm trong đơn xin trước đó hay không.	- 0: Khách hàng không có bảo hiểm trong đơn xin vay trước đó. Điều này có thể có nghĩa là khách hàng không mua bảo hiểm hoặc không đáp ứng được yêu cầu để được bảo hiểm. - 1: Khách hàng có bảo hiểm trong đơn xin vay trước đó. Điều này cho thấy khách hàng đã mua bảo hiểm hoặc đáp ứng được yêu cầu để được bảo hiểm.

Bảng 4.7. Bảng khám phá và mô tả các biến trong bảng dữ liệu "previous_application"

4.3.2.1. Kiểm tra và xem xét các giá trị rỗng

Index	COLUMN	Total	Percent
0	RATE_INTEREST_PRIVILEGED	1664263	99.643698
1	RATE_INTEREST_PRIMARY	1664263	99.643698
2	AMT_DOWN_PAYMENT	895844	53.636480
3	RATE_DOWN_PAYMENT	895844	53.636480
4	NAME_TYPE_SUITE	820405	49.119754
5	NFLAG_INSURED_ON_APPROVAL	673065	40.298129
6	DAYS_TERMINATION	673065	40.298129
7	DAYS_LAST_DUE	673065	40.298129
8	DAYS_LAST_DUE_1ST_VERSION	673065	40.298129
9	DAYS_FIRST_DUE	673065	40.298129

10	DAYS_FIRST_DRAWING	673065	40.298129
11	AMT_GOODS_PRICE	385515	23.081773
12	AMT_ANNUITY	372235	22.286665
13	CNT_PAYMENT	372230	22.286366
14	PRODUCT_COMBINATION	346	0.020716
15	AMT_CREDIT	1	0.000060

Bảng 4.8. Bảng mô tả tỷ lệ missing data theo từng thuộc tính trong bảng "previous_application"

Thuộc tính **"RATE_INTEREST_PRIVILEGED", "RATE_INTEREST_PRIMARY"** đều có tỷ lệ phần trăm giá trị missing xấp xỉ 99,64%. Tất cả các khoản vay có giá trị ở hai cột này để thuộc danh mục vay "POS". Hầu hết các đơn xin vay thuộc danh mục "POS" (99,14%) đều có giá trị ở hai cột này. Thông thường, trong các giao dịch POS, lãi suất và điều kiện vay được xác định trước và áp dụng cho tất cả các khách hàng có cùng loại giao dịch. Điều này có ý nghĩa rằng các khách hàng trong danh mục POS sẽ chia sẻ các điều kiện vay tương tự và không có sự chênh lệch quá lớn giữa các khách hàng.

Thuộc tính **"AMT_DOWN_PAYMENT"** (Số tiền trả trước cho đơn trước đó) có tỷ lệ phần trăm missing value xấp xỉ 56,64%. Lý do có thể xuất phát từ đặc điểm của khoản vay. Một số khoản vay có thể không yêu cầu trả trước hoặc không yêu cầu thông tin về số tiền trả trước. Trong một số trường hợp, các khoản vay như vay tiêu dùng hoặc vay tiền mặt có thể không yêu cầu số tiền trả trước từ khách hàng. Ví dụ, hầu hết các khoản vay tiêu dùng thuộc danh mục POS (99,99%), đều có giá trị ở cột "RATE_INTEREST_PRIMARY". Điều này có thể do các đơn vay POS thường liên quan đến việc mua sắm tại điểm bán hàng hoặc trả góp mua hàng, và do đó có yêu cầu

lãi suất cơ bản cụ thể. Danh mục "POS" trong bộ dữ liệu Home Credit thường áp dụng cho các giao dịch trực tiếp tại các cửa hàng hoặc điểm bán hàng, trong đó người vay có thể mua sắm và trả góp sản phẩm hoặc dịch vụ. Điều này đặc biệt phổ biến trong các ngành như điện tử, điện thoại di động, đồ gia dụng, và các sản phẩm tiêu dùng khác. Vì các đơn xin vay POS liên quan đến việc mua sắm và trả góp, thông tin về lãi suất cơ bản (AMT_DOWN_PAYMENT) thường được cung cấp để người vay có thể biết được mức lãi suất áp dụng cho khoản vay. Do đó, đa số các đơn xin vay thuộc danh mục "POS" trong bộ dữ liệu Home Credit thường có giá trị trong cột "AMT_DOWN_PAYMENT". Trong khi đó, đối với các đơn xin vay liên quan đến thẻ tín dụng “Cards”, (AMT_DOWN_PAYMENT) có thể không được cung cấp. Chỉ có 0,002% đơn xin vay trong danh mục “Cards” có thông tin về lãi suất thông tin về lãi suất cơ bản. Điều này có thể do các thẻ tín dụng có thể áp dụng các hình thức tính lãi suất khác nhau, chẳng hạn như lãi suất hằng tháng hoặc lãi suất không áp dụng. Do đó, cột "AMT_DOWN_PAYMENT" có thể rỗng trong trường hợp này.

Thuộc tính **“RATE_DOWN_PAYMENT”** (Tỷ lệ trả trước được chuẩn hóa trên tín dụng trước đó), tương ứng với thuộc tính “AMT_DOWN_PAYMENT” (Số tiền trả trước cho đơn trước đó), do đó, có phần trăm missing value xấp xỉ 56,64%. Lý do có thể xuất phát từ đặc điểm của khoản vay.

Thuộc tính **“NAME_TYPE_SUITE”** có phần trăm giá trị missing xấp xỉ 49,12%. Điều này có thể xuất phát từ các lý do khách quan như khách hàng không có người đi cùng, khách hàng không cung cấp thông tin này,... Tuy nhiên, thuộc tính này có thể đánh giá là không có nhiều ảnh hưởng lớn khi xét đến khả năng thanh toán tín dụng của khách hàng bên cạnh các yếu tố chính như thu nhập, lịch sử tín dụng, tình hình công việc,...của khách hàng.

Thuộc tính **“NFLAG_INSURED_ON_APPROVAL”** có phần trăm giá trị missing xấp xỉ 40,3%. Có thể có nhiều lý do dẫn đến việc giá trị rỗng, bao gồm việc thiếu thông tin, lỗi ghi nhận dữ liệu, hoặc khách hàng không có thông tin bảo hiểm,... Tuy nhiên, việc có được bảo hiểm không đảm bảo rằng khách hàng sẽ có khả năng thanh toán tín dụng tốt. Vì vậy, trong quá trình phân tích khả năng tín dụng, yếu tố có được bảo hiểm chỉ là một trong nhiều yếu tố cần xem xét.

TP.HCM, ngày 15 tháng 6 năm 2023

Thuộc tính **"DAYS_FIRST_DRAWING", "DAYS_FIRST_DUE", "DAYS_LAST_DUE_1ST_VERSION", "DAYS_LAST_DUE"**, và **"DAYS_TERMINATION"** có tỷ lệ giá trị missing là 40,3%. Lý do là các giá trị này thuộc vào các khoản vay đã bị hủy, bị từ chối, hoặc không được sử dụng như đã đề xuất. Tất cả các khoản vay này đều có giá trị rỗng ở các cột "DAYS_FIRST_DRAWING", "DAYS_FIRST_DUE", "DAYS_LAST_DUE_1ST_VERSION", "DAYS_LAST_DUE", và "DAYS_TERMINATION". Lý do là vì khi đơn xin vay đã bị hủy, bị từ chối, hoặc không được sử dụng như đã đề xuất thông tin về các ngày này thường không được cung cấp, ghi nhận. Tuy nhiên, vẫn có một số khoản vay được chấp thuận cũng có giá trị rỗng ở các cột này.

Thuộc tính **"AMT_GOODS_PRICE"** có tỷ lệ giá trị missing là 22,29%. Các giá trị này thuộc vào các khoản vay đã bị hủy, bị từ chối, hoặc không được sử dụng như đã đề xuất. Hầu hết các khoản vay này đều có giá trị rỗng ở cột "AMT_GOODS_PRICE". Tuy nhiên, vẫn có một số khoản vay được chấp thuận cũng có giá trị rỗng ở các cột này. Các khoản vay này thuộc hợp đồng sản phẩm Khoản vay "Revolving loans". Điều này cho thấy, một số khoản vay "Revolving loans" có thể không có giá trị trong cột "AMT_GOODS_PRICE". Vì cột "AMT_GOODS_PRICE" thường được sử dụng để ghi nhận giá trị hàng hóa, dịch vụ hoặc tài sản được mua bằng khoản vay. Tuy nhiên, với loại hợp đồng "Khoản vay Revolving loans", không có mục tiêu cụ thể để mua hàng hoá hoặc dịch vụ, mà thay vào đó người vay có quyền sử dụng số tiền vay một cách linh hoạt. Do đó, không có giá trị cụ thể để ghi nhận trong cột này hay đơn xin vay thuộc hợp đồng này không cần ghi nhận giá trị ở cột "AMT_GOODS_PRICE".

Thuộc tính **"AMT_ANNUITY"** có tỷ lệ giá trị missing là 22,29%. Các giá trị này đều thuộc vào các khoản vay đã bị hủy, bị từ chối, hoặc không được sử dụng như đã đề xuất. Lý do có thể là vì các khoản vay này không được hoàn tất và không có thông tin đầy đủ về kế hoạch trả nợ hàng tháng.

Thuộc tính **"CNT_PAYMENT"** có tỷ lệ giá trị missing là 22,29%. Lý do là các giá trị này đều thuộc vào các khoản vay đã bị hủy, bị từ chối, hoặc không được sử dụng

như đã đề xuất. Lý do có thể là vì các khoản vay này không được hoàn tất và không có thông tin đầy đủ về kỳ hạn của tín dụng.

Thuộc tính **“PRODUCT_COMBINATION”, “AMT_CREDIT”** có tỉ lệ phần giá trị missing thấp, lần lượt là 0.021% và 0.000060%. Do đó, có thể xem xét loại bỏ các dòng dữ liệu chứa giá trị missing trong hai cột này.

4.3.2.2. Phân tích các biến số

a. Thống kê mô tả

	AMT_ANNUITY	**AMT_APPLICATION**	**AMT_CREDIT**	**AMT_DOWN_PAYMENT**	**AMT_GOODS_PRICE**	**RATE_INTEREST_PRIMARY**	**RATE_INTEREST_PRIVILEGED**	**CNT_PAYMENT**
count	1297979.0	1670214.0	1670213.0	774370.0	1284699.0	5951.0	5951.0	1297984.0
mean	15955.12	175233.86	196114.02	6697.40	227847.28	0.19	0.77	16.05
std	14782.14	292779.76	318574.62	20921.5	315396.56	0.09	0.1	14.57

min	0.0	0.0	0.0	-0.9	0.0	0.03	0.37	0.0
25%	6321.78	18720.0	24160.5	0.0	50841.0	0.16	0.72	6.0
50%	11250.0	71046.0	80541.0	1638.0	112320.0	0.2	0.84	12.0
75%	20658.42	180360.0	216418.5	7740.0	234000.0	0.19	0.85	24.0
max	418058.15	6905160.0	6905160.0	3060045.0	6905160.0	1.0	1.0	84.0

Bảng 4.9. Bảng thống kê mô tả các biến trong bảng dữ liệu "previous_application"

Thuộc tính AMT_ANNUITY (Khoản trả góp hàng tháng của khoản vay trước):

Giá trị nhỏ nhất là 0.0 cho thấy có khách hàng không có số tiền trả hàng tháng hoặc trả 0 đồng. Đây có thể là đại diện cho các khoản vay thuộc các đơn xin vay đã bị hủy, bị từ chối hoặc không được sử dụng như đề xuất hoặc là các khoản vay thuộc danh mục sản phẩm thẻ tín dụng Cards.

Giá trị lớn nhất là 418,058.15, thể hiện số tiền trả hàng tháng lớn nhất là rất cao. Điều này có thể chỉ ra khả năng thanh toán tốt hơn với số tiền trả hàng tháng cao hơn.

Với tập khách hàng lớn, độ lệch chuẩn của khoản trả góp hàng tháng cao (14782.14) gần với giá trị trung bình của khoản trả góp hàng tháng (15955.12). Điều này cho thấy rằng dữ liệu trả góp hàng tháng có một mức độ biến động lớn, và có sự chênh lệch lớn giữa các giá trị trong tập dữ liệu.

Thuộc tính AMT_APPLICATION (Số tiền tín dụng mà khách hàng yêu cầu trong đơn xin vay trước):

Giá trị nhỏ nhất là 0.0 chỉ ra các yêu cầu khoản vay các đơn xin vay đã bị hủy, bị từ chối hoặc không được sử dụng như đề xuất. hoặc yêu cầu khoản vay với giá trị tài sản là 0.

Giá trị lớn nhất là 6905160.0 cho thấy có khách hàng yêu cầu khoản vay với giá trị tài sản lớn nhất là rất cao. Điều này có thể chỉ ra khả năng tài chính và khả năng thanh toán tốt.

Giá trị Mean là 175233.86, là giá trị trung bình của khoản vay mà khách hàng yêu cầu. Giá trị trung bình lớn, thuộc vào khoảng tứ phân vị thứ 3, điều này cho thấy rằng có sự tồn tại của khách hàng vay lớn trong tập dữ liệu.

Thuộc tính AMT_CREDIT (Số tiền tín dụng cuối cùng được chấp nhận trong đơn xin trước):

Tương ứng AMT_APPLICATION và AMT_DOWN_PAYMENT, giá trị nhỏ nhất, giá trị lớn nhất và giá trị Mean của AMT_CREDIT có ý nghĩa tương tự. Tuy nhiên, số tiền cấp AMT_CREDIT cao hơn so với số tiền yêu cầu AMT_APPLICATION, có thể cho thấy Home Credit đánh giá tốt khả năng tín dụng của khách hàng trong khả năng trả nợ.

Có thể so sánh AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_DOWN_PAYMENT với các yếu tố khác, đặc biệt là thu nhập hàng tháng của khách hàng, để có thể đánh giá khả năng thanh toán tín dụng. Nếu số tiền trả hàng tháng gấp đôi hoặc vượt quá thu nhập hàng tháng, có thể cho thấy khách hàng có thể gặp khó khăn trong việc thanh toán và có nguy cơ không đảm bảo khả năng tín dụng tốt.

Thuộc tính AMT_GOODS_PRICE (Giá hàng hóa mà khách hàng yêu cầu vay (nếu có) trong đơn vay trước đó):

Giá trị "AMT_GOODS_PRICE" bằng giá trị AMT_APPLICATION tương ứng:

Giá trị nhỏ nhất là 0, cho thấy có các đơn vay đã bị hủy, bị từ chối hoặc không được sử dụng như đề xuất. Hoặc giá trị hàng hóa, sản phẩm không được xác định hoặc không có giá trị được khai báo.

Giá trị AMT_GOODS_PRICE" lớn nhất là 6,905,160, là giá trị lớn nhất của giá trị hàng hóa, sản phẩm trong các đơn vay. Việc có giá trị hàng hóa cao như vậy có thể cho thấy khách hàng đang vay để mua các tài sản có giá trị cao. Khách hàng có khả năng tài chính tốt hơn và sẵn sàng vay số tiền lớn để mua các tài sản cao cấp.

Thuộc tính RATE_INTEREST_PRIMARY (Lãi suất cơ bản được chuẩn hóa trên tín dụng trước đó):

Giá trị nhỏ nhất cho biết mức lãi suất thấp nhất trong các đơn vay. Nếu giá trị nhỏ nhất là 0.03, điều này có thể cho thấy có các đơn vay với lãi suất thấp.

Giá trị lớn nhất thể hiện mức lãi suất cao nhất trong các đơn vay. Nếu giá trị Max là 1.0, điều này có thể chỉ ra tỷ lệ lãi suất tối đa mà khách hàng phải trả. Mức lãi suất cao có thể tạo áp lực lên khả năng thanh toán và ảnh hưởng đến khả năng tín dụng của khách hàng.

Giá trị trung bình của lãi suất trong các đơn vay là 0.19, điều này cho thấy lãi suất trung bình mà khách hàng phải trả cho khoản vay. Giá trị này có thể so sánh với lãi suất trung bình trên thị trường hoặc với lãi suất được áp dụng trong ngành công nghiệp tài chính để đánh giá tính hợp lý và khả năng thanh toán của khách hàng.

Thuộc tính RATE_INTEREST_PRIVILEGED (Lãi suất ưu đãi được chuẩn hóa trên tín dụng trước đó):

Giá trị nhỏ nhất là 0.37, cho biết tỷ lệ lãi suất ưu đãi thấp nhất trong các đơn vay. Điều này có thể chỉ ra rằng một số khách hàng đã có lợi thế về lãi suất, có mối quan hệ lâu dài với ngân hàng, có lịch sử tín dụng tốt và thể hiện khả năng thanh toán đáng tin cậy.

Giá trị lớn nhất là 1.0 là tỷ lệ lãi suất ưu đãi cao nhất trong các đơn vay. Giá trị này có thể chỉ ra rằng khách hàng có khả năng nhận được lãi suất ưu đãi cao hơn như

vậy là do có điều kiện tài chính tốt hơn. Và để được nhận lãi suất ưu đãi, khách hàng có thể cần đáp ứng một số yêu cầu như thu nhập cao, lịch sử tín dụng tốt, hoặc tài sản đảm bảo để đảm bảo khả năng trả nợ.

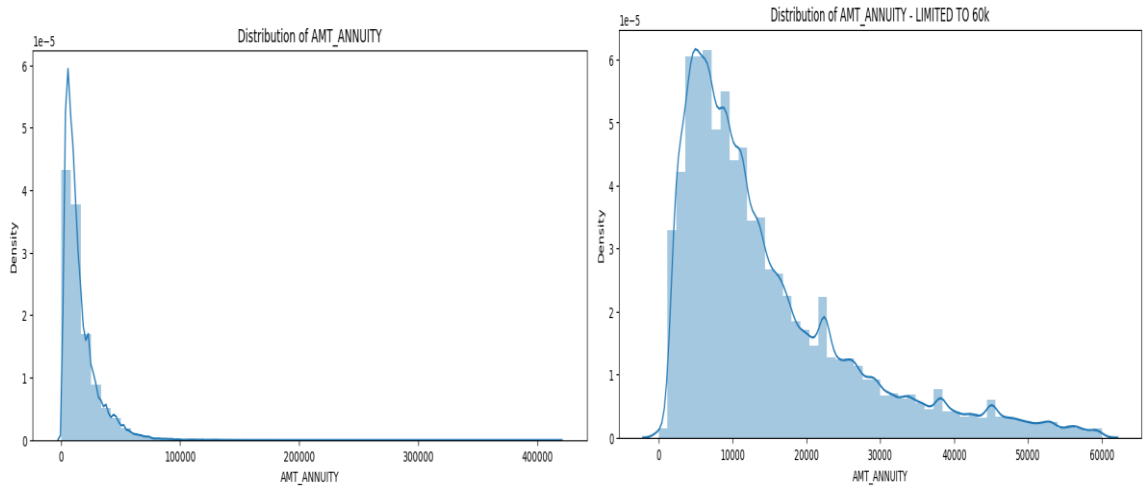
Thuộc tính CNT_PAYMENT (Kỳ hạn (số lần thanh toán) của tín dụng trước đó tại thời điểm đơn xin trước đó):

Giá trị nhỏ nhất của CNT_PAYMENT là 0, cho thấy có các đơn vay đã bị hủy, bị từ chối hoặc không được sử dụng như đề xuất. Bên cạnh đó.

Giá trị lớn nhất của CNT_PAYMENT cho biết số lần trả nợ tối đa trong thời hạn vay. Điều này cho thấy một số khoản vay có thời gian trả nợ kéo dài và yêu cầu nhiều lần trả nợ hàng tháng trong suốt thời gian vay.

Giá trị trung bình của CNT_PAYMENT là 16.05, đây là số lần trung bình khách hàng phải trả nợ hàng tháng trong thời hạn vay. Giá trị này có thể thể hiện mức độ tải trọng tài chính của khách hàng, tức là số lần phải trả nợ hàng tháng.

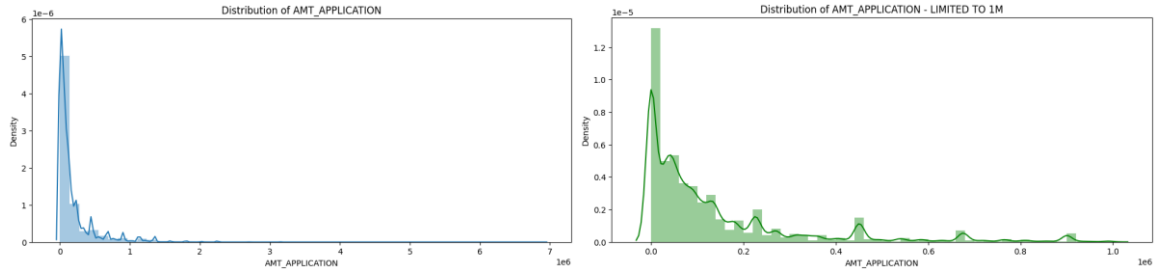
b. Phân phối dữ liệu



Hình 4.14. Biểu đồ phân phối của thuộc tính "AMT_ANNUITY"

(Nguồn: Nhóm tác giả)

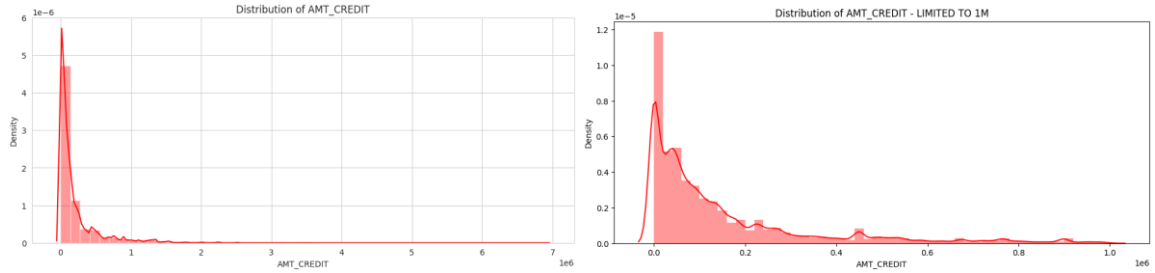
Qua biểu đồ trên, chúng ta nhận thấy rằng số tiền trả hàng tháng (AMT_ANNUITY) trong các đơn đăng ký vay trước được phân bố rộng và tập trung chủ yếu trong khoảng giá trị từ 10.000 đến 20.000. Điều này cho thấy số tiền trả hàng tháng thông thường của các khoản vay trong tập dữ liệu nằm trong khoảng này.



Hình 4.15. Biểu đồ phân phối của thuộc tính "AMT_APPLICATION"

(Nguồn: Nhóm tác giả)

Biểu đồ trên cho biết phân phối của khoản tiền vay khách hàng yêu cầu trong các đơn đăng ký vay trước (AMT_APPLICATION). Biểu đồ cho thấy sự đa dạng trong các khoản vay và tập trung chủ yếu trong khoảng giá trị dưới 200,000. Điều này có nghĩa là số lượng đơn đăng ký vay trước với khoản vay nhỏ hơn hoặc bằng 200,000 chiếm tỷ lệ cao hơn so với các khoản vay lớn hơn. Biểu đồ cũng cho thấy rằng phân phối của khoản vay có đuôi dài, tức là vẫn có một số đơn đăng ký vay trước với các khoản vay lớn hơn 200,000, nhưng tần suất của chúng ít hơn so với khoản vay nhỏ hơn.



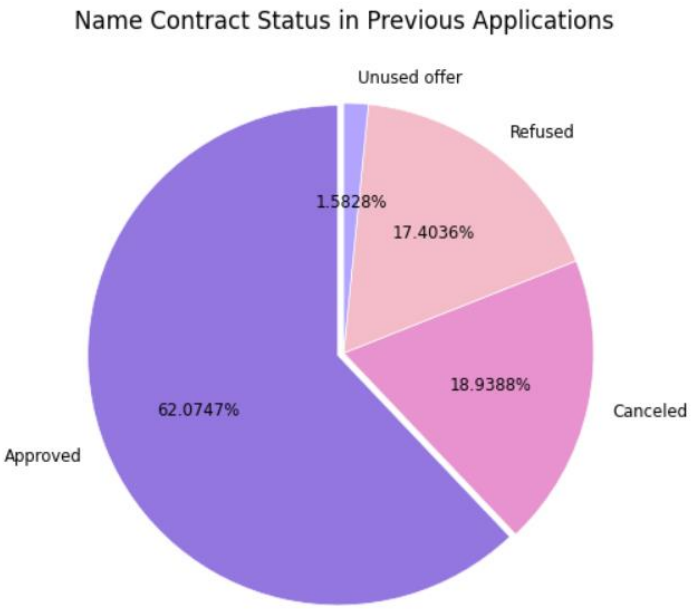
Hình 4.16. Biểu đồ phân phối của thuộc tính "AMT_CREDIT"

(Nguồn: Nhóm tác giả)

Tương tự (AMT_APPLICATION), biểu đồ phân phối (distplot) của khoản vay (AMT_CREDIT) trong các đơn đăng ký vay trước cho thấy sự đa dạng và phân bố rộng của các khoản vay được Home Credit cung cấp. Điểm chính là tập trung chủ yếu của các khoản vay nằm trong khoảng giá trị dưới 200,000. Điều này cho thấy Home Credit có sự linh hoạt trong việc cung cấp các khoản vay với mức giá đa dạng để đáp ứng nhu cầu vay của khách hàng. Các khoản vay có giá trị dưới 200,000 được tập trung nhiều hơn, có thể phản ánh việc Home Credit hướng đến khách hàng có nhu cầu vay

nhỏ hơn và tạo điều kiện thuận lợi cho việc vay vốn trong khoản này. Tuy nhiên, biểu đồ cũng cho thấy có một số khoản vay lớn hơn 200,000, mặc dù tần suất của chúng thấp hơn. Điều này cho thấy Home Credit cũng cung cấp các khoản vay lớn hơn cho khách hàng có nhu cầu tài chính cao hơn.

4.3.2.3. Phân tích các biến phân loại



Hình 4.17. Biểu đồ thể hiện phần trăm các trạng thái của hợp đồng của các đơn vay trước

(Nguồn: Nhóm tác giả)

Biểu đồ cho thấy tỷ lệ phần trăm của các loại trạng thái hợp đồng của đơn xin vay trước:

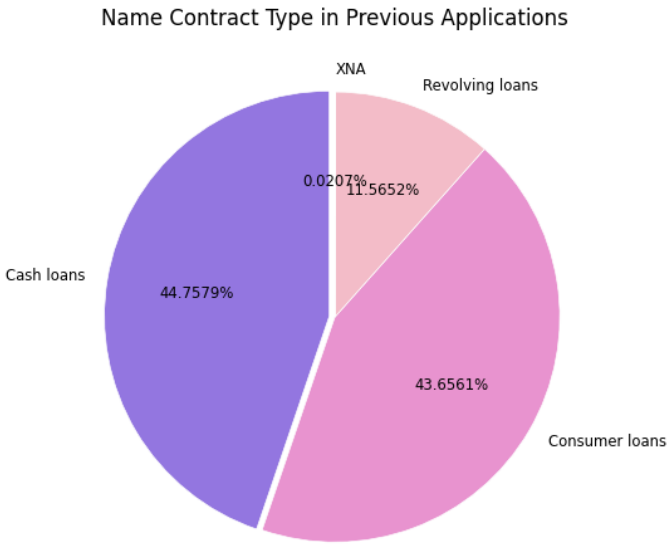
+ Tỉ lệ chấp thuận các đơn xin vay trước chiếm 62,074%, đứng đầu trong số các loại trạng thái hợp đồng. Điều này cho thấy tỷ lệ chấp thuận đơn xin vay trước đó của Home Credit là khá cao.

+ Tỷ lệ hủy bỏ đơn xin vay (19.5%) cũng khá đáng chú ý. Có thể tìm hiểu lý do tại sao khách hàng lại hủy bỏ đơn xin vay.

+ Tỷ lệ đơn xin vay bị từ chối (17.9%) là không nhỏ. Có thể chỉ ra rằng Home Credit có tiêu chí và quy trình chấp thuận khá nghiêm ngặt. Khách hàng có thể gặp

khó khăn trong việc đáp ứng các yêu cầu hoặc không đáp ứng được các tiêu chí cần thiết để được chấp thuận.

+ Tỷ lệ đơn xin vay với ưu đãi không sử dụng (1.6%) là thấp nhất. Điều này cho thấy số lượng đơn xin vay với ưu đãi không được sử dụng là khá ít. Các khách hàng đều tận dụng và chấp nhận các ưu đãi được cung cấp trong quá trình xin vay.



Hình 4.18. Biểu đồ thể hiện phần trăm các loại của hợp đồng của các đơn vay trước

(Nguồn: Nhóm tác giả)

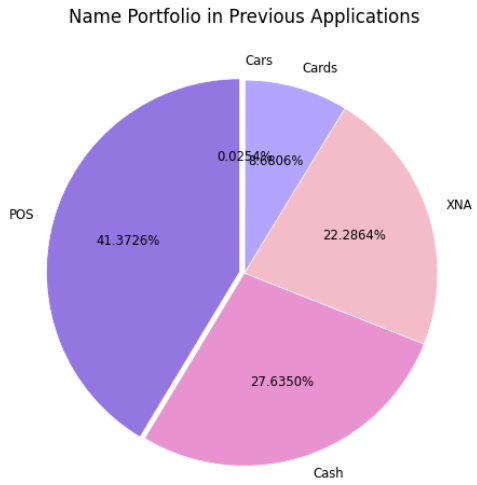
Biểu đồ cho thấy tỷ lệ phần trăm của các loại hợp đồng sản phẩm của đơn xin vay trước:

+ Cash loans chiếm tỷ lệ cao nhất trong các đơn xin vay trước đó. Điều này cho thấy nhu cầu vay tiền mặt của khách hàng khá phổ biến.

+ Consumer loans là loại hợp đồng vay tiền thứ hai phổ biến sau Cash loans. Điều này cho thấy nhu cầu vay tiền cho mục đích tiêu dùng của khách hàng cũng rất cao.

+ Revolving loans có tỷ lệ thấp hơn so với các loại hợp đồng khác. Tuy tỷ lệ này thấp, nhưng vẫn có một số khách hàng có nhu cầu sử dụng hình thức vay này.

+ XNA có tỷ lệ phần trăm rất thấp, gần như không đáng kể. Điều này có thể chỉ ra rằng có rất ít đơn xin vay trước đó mà không rõ hoặc không xác định được thông tin về trạng thái.



Hình 4.19. Biểu đồ thể hiện phần trăm các loại danh mục của hợp đồng của các đơn vay trước

(Nguồn: Nhóm tác giả)

Biểu đồ cho thấy tỷ lệ phần trăm của các loại danh mục của đơn xin vay trước:

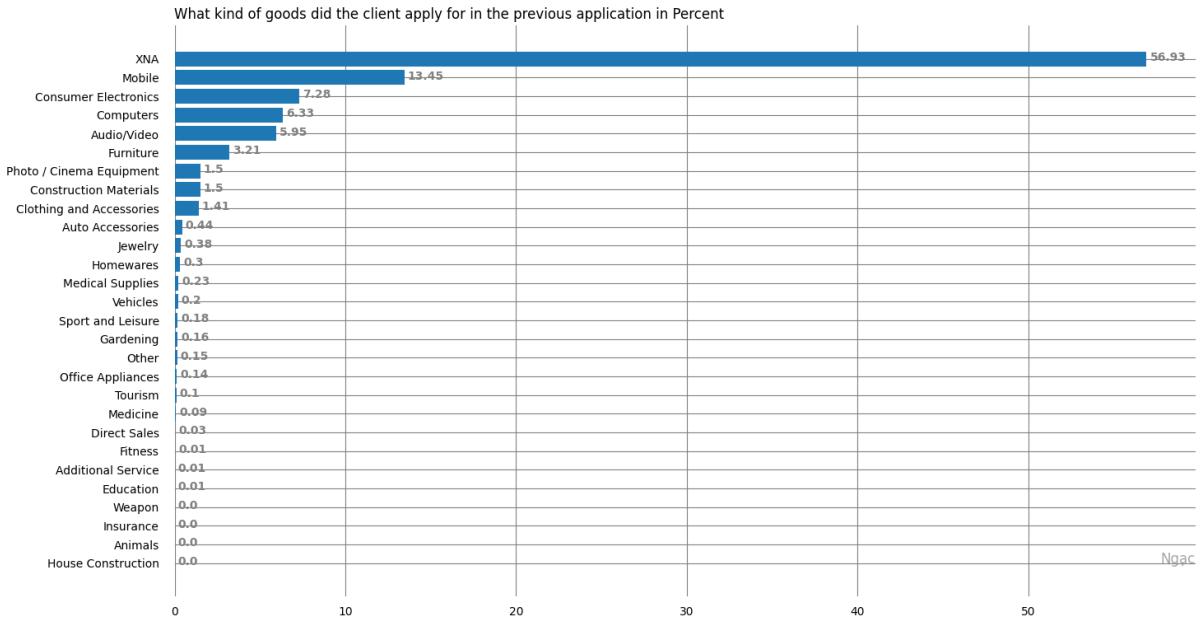
+ Giao dịch POS chiếm tỷ lệ cao nhất trong các giao dịch. Điều này cho thấy sử dụng các điểm bán hàng (Point of Sale) để thanh toán hoặc giao dịch là phổ biến và được sử dụng rộng rãi.

+ Giao dịch tiền mặt chiếm tỷ lệ cao trong các giao dịch. Điều này cho thấy một số khách hàng vẫn ưa thích giao dịch và thanh toán bằng tiền mặt.

+ Giao dịch có trạng thái XNA chiếm tỷ lệ khá cao. Tuy nhiên, không có đủ thông tin để xác định rõ ràng ý nghĩa của trạng thái này.

+ Giao dịch sử dụng thẻ (Cards) chiếm tỷ lệ khá thấp trong các giao dịch. Điều này có thể cho thấy việc sử dụng thẻ thanh toán không phổ biến hoặc các giao dịch khác không yêu cầu sử dụng thẻ thanh toán.

+ Giao dịch liên quan đến xe ô tô (Cars) chiếm tỷ lệ rất thấp. Điều này cho thấy việc vay tiền hoặc giao dịch liên quan đến ô tô không phổ biến.



Hình 4.20. Biểu đồ thể hiện phần trăm các loại của sản phẩm hợp đồng của các đơn vay trước

(Nguồn: Nhóm tác giả)

Biểu đồ cho thấy các sản phẩm mà khách hàng cung cấp trong đơn xin vay.

+ "XNA": Chiếm tỷ lệ lớn nhất trong bảng dữ liệu, với khoảng 56.93%. Tuy nhiên, điều đáng chú ý là không có thông tin cụ thể về ý nghĩa của trạng thái này.

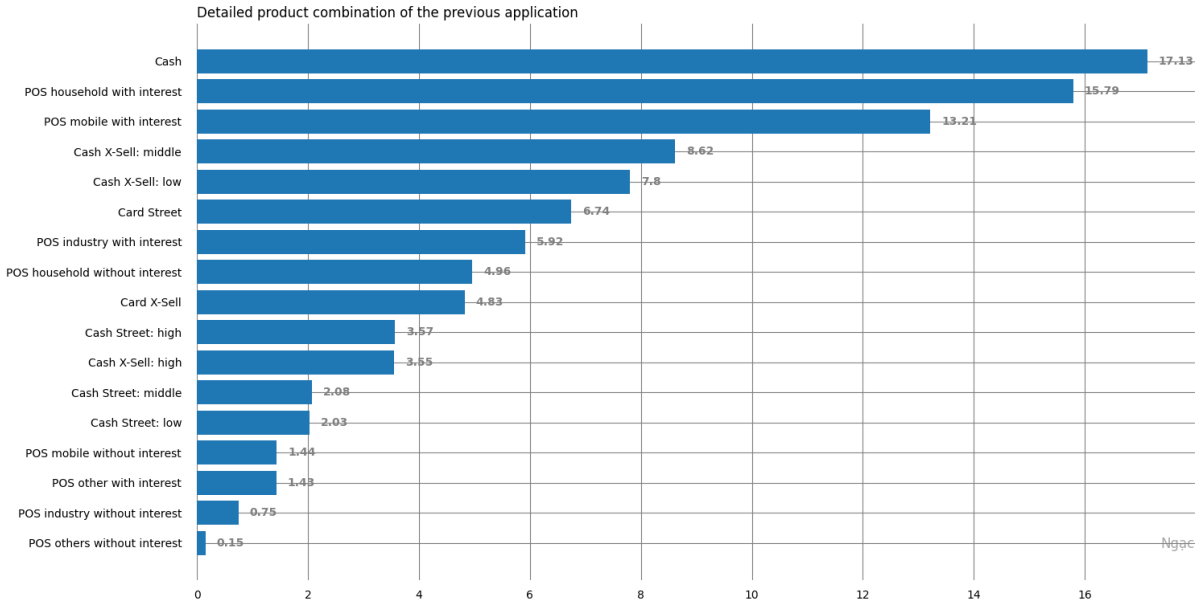
+ Giao dịch liên quan đến điện thoại di động (Mobiles) chiếm khoảng 13.45% trong bảng dữ liệu, tỷ lệ cao khách hàng đã thực hiện các giao dịch mua sắm hoặc vay vốn để mua điện thoại di động. Điều này có thể chỉ ra rằng điện thoại di động là một mặt hàng phổ biến và được ưa chuộng nhất (không tính đến giá trị "XNA" không xác định).

+ Giao dịch liên quan đến đồ điện tử tiêu dùng ("Consumer Electronics"): Loại giao dịch này chiếm khoảng 7.28% trong bảng dữ liệu. Điều này cho thấy một phần khá lớn khách hàng đã có nhu cầu mua các thiết bị điện tử tiêu dùng như máy ảnh, máy quay, hoặc các sản phẩm công nghệ khác.

+ Giao dịch liên quan đến máy tính ("Computers"): Loại giao dịch này chiếm khoảng 6.33% trong bảng dữ liệu. Điều này cho thấy một phần khá đáng kể khách hàng đã mua sắm hoặc vay vốn để mua máy tính, laptop hoặc các thiết bị liên quan.

+ Giao dịch liên quan đến nội thất ("Audio/Video"): Tỷ lệ 5.95% cho thấy có một phần nhỏ nhưng không phải là ít khách hàng quan tâm đến.

+ Giao dịch liên quan đến nội thất ("Furniture"): Tỷ lệ 3.21% cho thấy tỷ lệ khá nhỏ của khách hàng quan tâm đến sản phẩm này. Điều này có thể chỉ ra rằng không phải tất cả khách hàng đặt nhu cầu ưu tiên cho việc sử dụng sản phẩm này.



Hình 4.21. Biểu đồ thể hiện phần trăm các loại của sản phẩm lết hợp của các đơn vay trước

(Nguồn: Nhóm tác giả)

Biểu đồ cho thấy tỷ lệ phần trăm của các loại sản phẩm kết hợp của đơn xin vay trước. Và 7 loại hình sản phẩm kết hợp có tỉ lệ phần trăm cao nhất bao gồm:

+ Tiền mặt (Cash) chiếm tỷ lệ cao nhất với khoảng 17,13%. Điều này cho thấy một số lượng đáng kể khách hàng lựa chọn thanh toán bằng tiền mặt. Có thể rằng họ ưa thích phương thức thanh toán trực tiếp hoặc sử dụng các dịch vụ liên quan đến tiền mặt.

+ Giao dịch POS giao dịch sản phẩm gia dụng có lãi suất chiếm khoảng 15,79% trong bảng dữ liệu. Tỷ lệ này cho thấy một số khách hàng đáng kể đã thực hiện giao dịch liên quan đến mua sắm hoặc vay vốn cho các sản phẩm hoặc dịch vụ gia đình. Có thể rằng trong quá trình này, các giao dịch được áp dụng lãi suất. Điều này có thể chỉ

ra rằng khách hàng có nhu cầu đáng kể trong việc mua sắm các sản phẩm hoặc dịch vụ gia đình và sẵn sàng chấp nhận việc trả lãi suất để thực hiện giao dịch.

+ Giao dịch POS giao dịch điện thoại di động có lãi suất (POS mobile with interest): Chiếm khoảng 13,21%. Điều này cho thấy một phần khách hàng đã thực hiện giao dịch liên quan đến mua sắm hoặc vay vốn để mua điện thoại di động. Lãi suất được áp dụng có thể là do mức trả góp hoặc hình thức vay mượn.

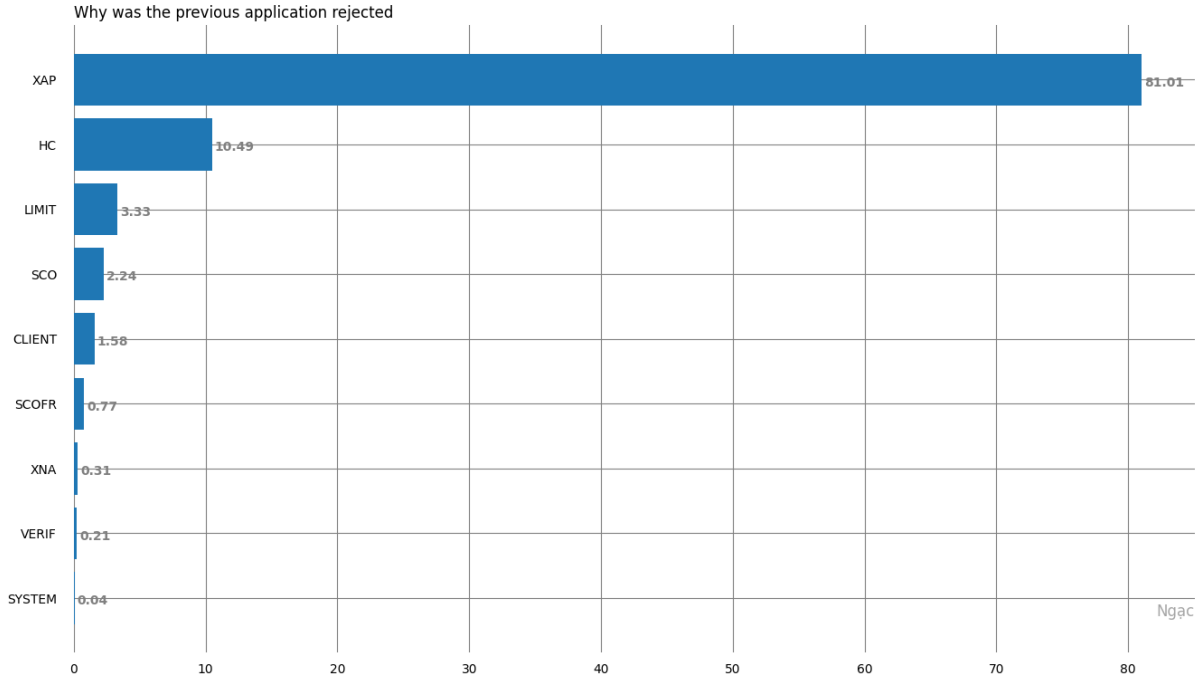
+ Tiền mặt giao dịch X-Sell: trung (Cash X-Sell: middle) chiếm tỷ lệ 8,62% trong bảng dữ liệu. Đây là một danh mục giao dịch tiền mặt có tỷ lệ cao, cho thấy một số khách hàng đã nhận được đề xuất về các giao dịch X-Sell liên quan đến tiền mặt với mức trung bình. Điều này có thể đồng nghĩa với việc các khách hàng trong danh mục này đã được đề xuất giao dịch bổ sung liên quan đến tiền mặt, có thể là các ưu đãi đặc biệt hoặc dịch vụ đi kèm.

+ Tiền mặt X-Sell: thấp (Cash X-Sell: low) chiếm khoảng 7,8% trong bảng dữ liệu. Tỷ lệ này cho thấy một phần khách hàng đã nhận được đề xuất giao dịch X-Sell liên quan đến tiền mặt với mức thấp. Điều này có thể ám chỉ đến các giao dịch có giá trị thấp hơn hoặc ít ưu đãi hơn so với danh mục trung. Thông qua việc đề xuất giao dịch tiền mặt X-Sell với mức thấp, có thể tạo điều kiện thuận lợi cho khách hàng thực hiện các giao dịch nhỏ hơn hoặc có giá trị không quá cao. Điều này có thể là các giao dịch hàng ngày, ví dụ như mua sắm nhỏ, thanh toán hóa đơn hay các chi tiêu nhỏ khác.

+ Giao dịch thẻ tín dụng (Card Street) chiếm khoảng 6,74% trong bảng dữ liệu. Tỷ lệ này cho thấy một số khách hàng đã thực hiện giao dịch bằng thẻ tín dụng tại các địa điểm giao dịch trực tiếp, như các giao dịch mua sắm hoặc thanh toán hàng ngày sử dụng thẻ tín dụng. Lý do có thể xuất phát từ việc sử dụng thẻ tín dụng để thực hiện giao dịch tại các địa điểm giao dịch trực tiếp cung cấp sự tiện lợi và linh hoạt cho khách hàng.

+ Giao dịch POS ngành công nghiệp có lãi suất (POS industry with interest) chiếm khoảng 5,92% trong bảng dữ liệu. Tỷ lệ này cho thấy một số khách hàng đã thực hiện giao dịch liên quan đến mua sắm hoặc vay vốn trong ngành công nghiệp. Lãi suất có thể liên quan đến việc mua sắm các sản phẩm hoặc dịch vụ từ các doanh nghiệp hoạt động trong ngành công nghiệp như sản xuất, xây dựng, vận chuyển, và các dịch

vụ công nghiệp khác. Lãi suất trong giao dịch này có thể liên quan đến mục đích như tạo ra lợi nhuận của doanh nghiệp, tăng giá trị giao dịch thông qua việc trả góp hoặc hình thức vay vốn, hoặc điều chỉnh giá trị thanh toán dựa trên thời gian sử dụng sản phẩm hoặc dịch vụ.



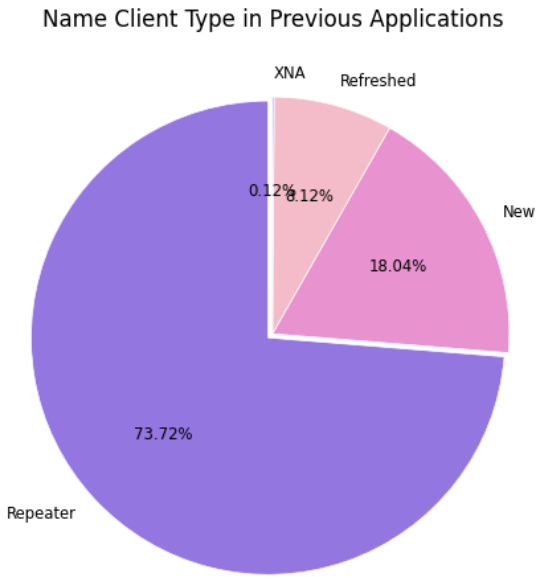
Hình 4.22. Biểu đồ thể hiện phần trăm các lý do các hợp đồng của các đơn vay trước bị từ chối

(Nguồn: Nhóm tác giả)

Biểu đồ trên cho biết lý do tại sao đơn đăng ký vay bị từ chối.

+ Giá trị "XAP" chiếm tỷ lệ phần trăm cao nhất 81,01%, đại diện cho các trường hợp mà nguyên nhân từ chối không được ghi lại hoặc không có thông tin cụ thể nào được cung cấp.

+ Ngoại trừ giá trị "XAP", lý do đơn đăng ký vay bị từ chối nhiều nhất (10,49%) là vì không đáp ứng các yêu cầu tín dụng cụ thể của Home Credit hoặc có các vấn đề về lịch sử tín dụng của khách hàng (được ký hiệu với mã "HC").



Hình 4.23. Biểu đồ thể hiện phần trăm các loại của khách hàng của các đơn vay trước

(Nguồn: Nhóm tác giả)

Biểu đồ trên cho biết tỷ lệ phần trăm các loại khách hàng đăng ký khoản vay.

+ Tỷ lệ khách hàng "Repeater (Reap)" này chiếm đa số với 1,231,261 trường hợp (80.24%). Điều này cho thấy rằng có nhiều khách hàng đã từng vay và tái vay trong quá khứ. Điều này có thể cho thấy khả năng tín dụng tốt và sự tin tưởng của ngân hàng đối với khách hàng này.

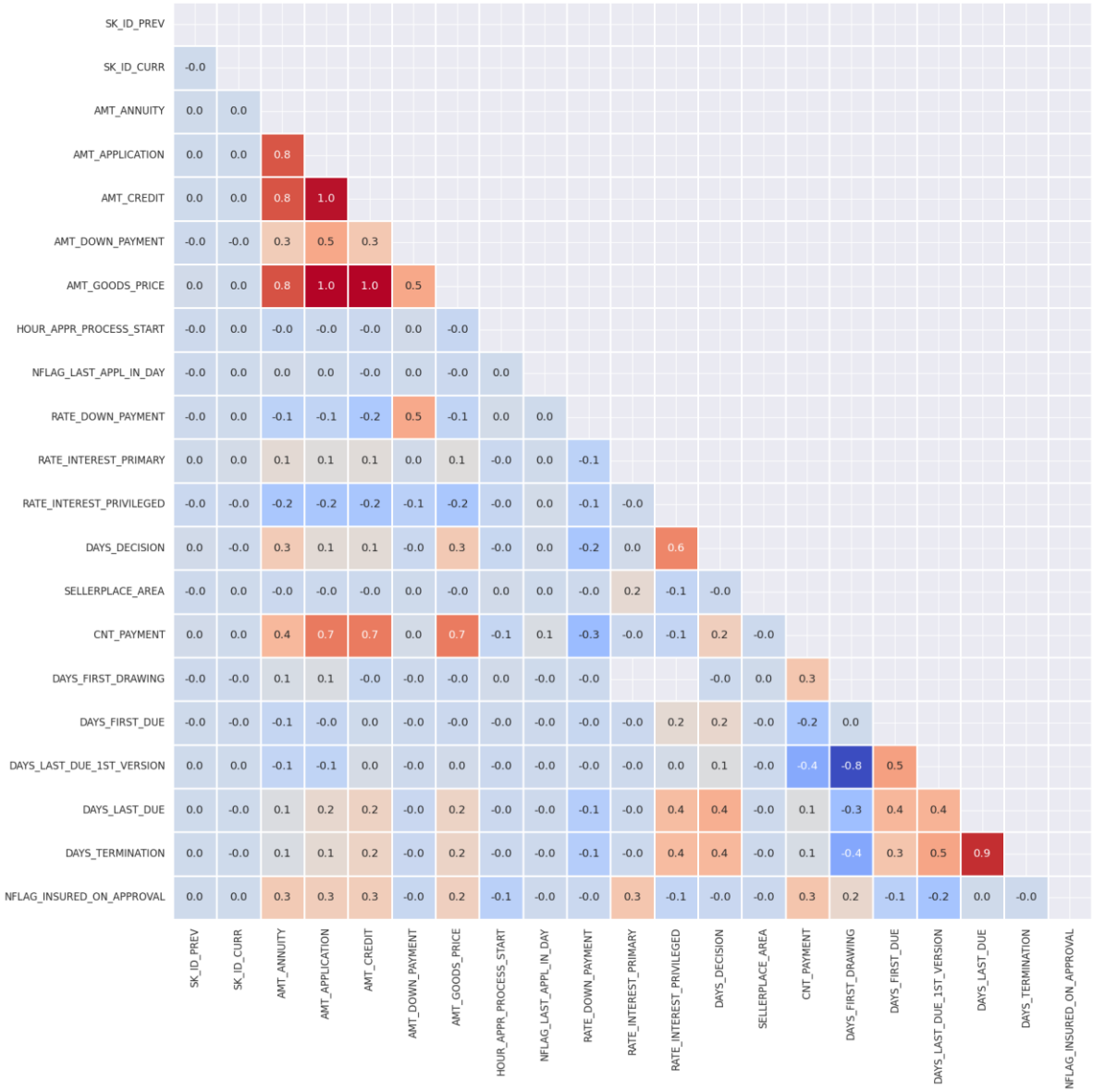
+ Tỉ lệ khách hàng "New" chiếm 301,363 trường hợp (19.68%). Đây là khách hàng mới và chưa từng vay trước đây. Việc có một số lượng đáng kể khách hàng mới có thể tương đồng về khả năng tín dụng của họ, và ngân hàng có thể đánh giá rủi ro và khả năng thanh toán của họ dựa trên thông tin và tiêu chí khác.

+ Các khách hàng Refreshed chiếm 135,649 trường hợp (8.87%). Đây là khách hàng đã từng vay trước đây và đang cần làm mới hồ sơ vay. Việc làm mới hồ sơ có thể liên quan đến việc gia hạn hoặc thay đổi điều khoản hợp đồng vay. Tỷ lệ này có thể cho thấy khách hàng đang duy trì một mức độ tín dụng ổn định và có nhu cầu tiếp tục vay.

+ XNA: Tỉ lệ này rất thấp với chỉ 1,941 trường hợp (0.13%). Đây là giá trị không xác định hoặc không xác định rõ nguyên nhân. Với tỷ lệ này thấp, không thể đưa ra các kết luận chính xác về khả năng tín dụng của khách hàng trong trường hợp này.

4.3.3. Tương quan giữa các biến

4.3.3.1. Hệ số tương quan

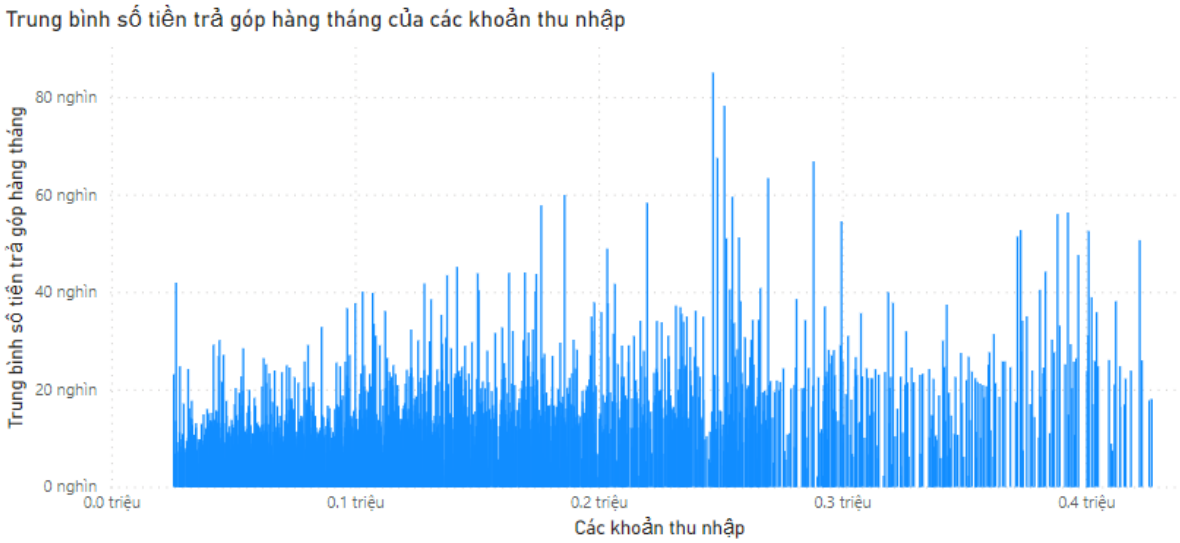


Hình 4.24. Biểu đồ thể hiện tương quan giữa các biến trong bảng "previous_application"

(Nguồn: Nhóm tác giả)

Hầu hết các thuộc tính đều có mối tương quan yếu với nhau. Các thuộc tính có mối tương quan mạnh bao gồm: “AMT_CREDIT” với “AMT_APPLICATION” và “AMT_GOODS_PRICE” (1.0); “DAYS_TERMINATION” với “DAYS_LAST_DUE” (0.9), “AMT_ANNUITY” với “AMT_CREDIT”, “AMT_APPLICATION” và “AMT_GOODS_PRICE” (0.8), “CNT_PAYMENT” với “AMT_APPLICATION”, “AMT_CEDIT”, “AMT_GOODS_PRICE” (0.7), cho thấy hai đặc trưng này giải thích dữ liệu một cách rất tương tự nhau, có thể dẫn đến hiện tượng overfitting cho mô hình. Cần xử lý tương quan đa cộng tuyến này bằng cách áp dụng các biện pháp như loại bỏ một trong hai đặc trưng. Bên cạnh đó, hầu hết các mô hình học máy đều giả định rằng không có đa cộng tuyến cao giữa các đặc trưng. Khi đa cộng tuyến cao xảy ra, mô hình có thể trở nên không ổn định và dự đoán không chính xác.

4.3.3.2. Trực quan hóa mối quan hệ giữa các biến

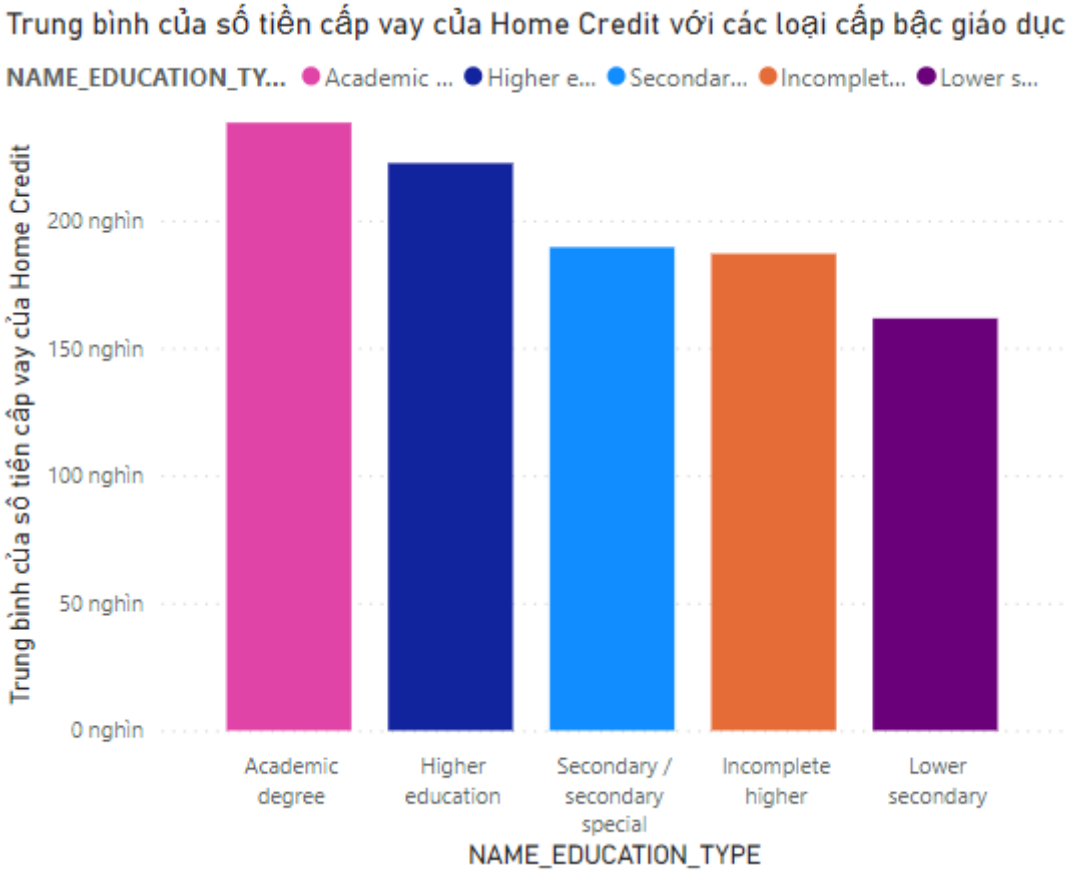


Hình 4.25. Biểu đồ thể hiện giá trị trung bình số tiền trả góp hàng tháng của các khoản thu nhập

(Nguồn: Nhóm tác giả)

Biểu đồ thể hiện số tiền trả góp trung bình hàng tháng của các khoản thu nhập dưới số tiền 427500 (giá trị thuộc tính “AMT_INCOME_TOTAL” tập trung phân phối ở khoản thu nhập này). Các khoản vay có mức thu nhập dưới 100,000 có mức trả góp

hàng tháng tương đối thấp, đều dưới 50,000. Điều này cho thấy rằng đa số khách hàng có khoản vay nhỏ và có thu nhập thấp đều có khả năng trả góp hàng tháng trong khoản tiền này. Số tiền trả góp lớn nhất có giá trị hơn 80000, lại nhiên lại nằm trong khoản thu nhập lớn hơn số tiến 20000. Nhóm khách hàng có khoản thu nhập nằm trong khoản này có khả năng trả nợ tốt, họ có khả năng tài chính ổn định và có thể đáp ứng các khoản trả góp hàng tháng lớn.

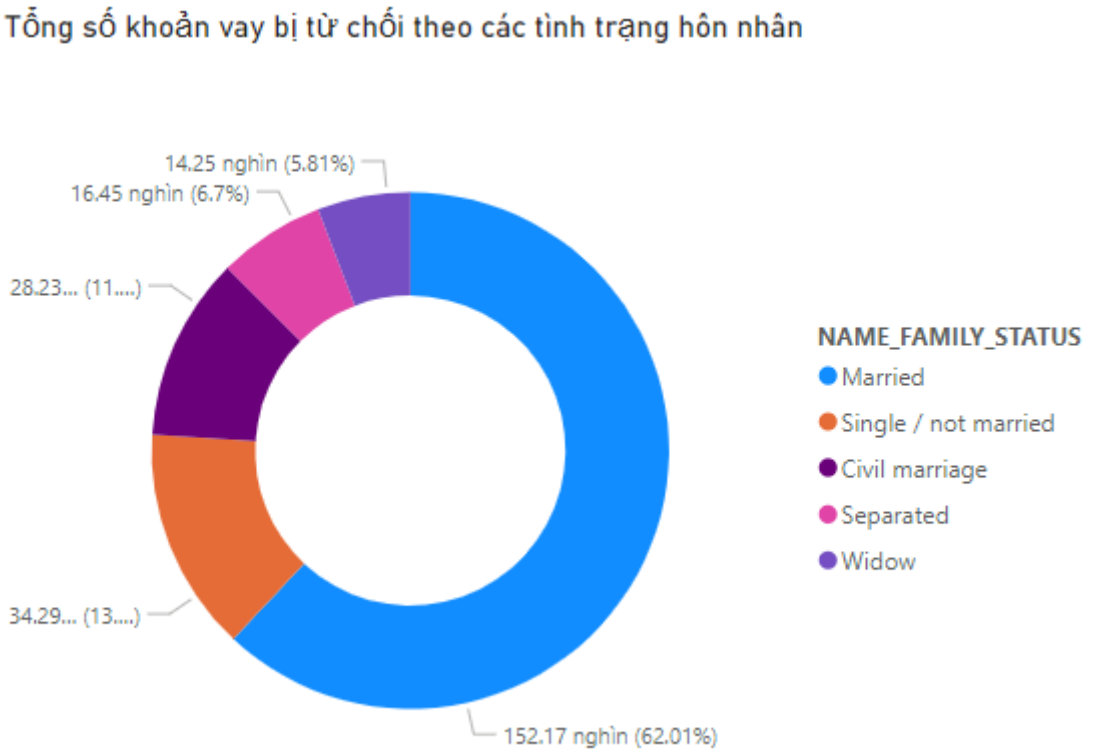


Hình 4.26. Biểu đồ thể hiện giá trị trung bình số tiền cấp vay của Home Credit với các loại cấp bậc giáo dục

(Nguồn: Nhóm tác giả)

Biểu đồ biểu diễn số tiền trung bình mà Home Credit cấp vay cho các loại cấp bậc giáo dục của khách hàng. Rõ ràng, các khoản vay mà khách hàng có loại cấp bậc giáo dục "Đã đạt được một bằng cấp học thuật cao hơn" được cấp khoản vay cao nhất. Điều này cho thấy Home Credit đánh giá cao sự ổn định tài chính và khả năng trả nợ của nhóm khách hàng này. Họ được xem là những người có khả năng trả nợ đáng tin

cậy. Tiếp đến là nhóm khách hàng có cấp bậc giáo dục là “Giáo dục đại học” với khoản được cấp vay gần bằng các khách hàng có loại cấp bậc giáo dục “Đã có bằng đại học”, thể hiện Home Credit vẫn đánh giá cao những người có trình độ giáo dục cao hơn so với nhóm có cấp bậc giáo dục thấp hơn. Nhóm “Trung học phổ thông” nhận được khoản vay thấp hơn so với nhóm Đã có bằng đại học và Giáo dục đại học, tuy nhiên, vẫn được cung cấp một mức vay đáng kể. Đối với các cấp bậc giáo dục khác, là nhóm nhận được khoản vay thấp nhất trong các loại cấp bậc giáo dục, có thể do yếu tố rủi ro tài chính cao hơn hoặc độ tin cậy thấp hơn của những người không thuộc các cấp bậc giáo dục chính.



Hình 4.27. Biểu đồ thể hiện tổng số khoản vay bị từ chối theo các tình trạng hôn nhân

(Nguồn: Nhóm tác giả)

Biểu đồ thể hiện số lượng khoản vay bị từ chối tính theo yếu tố tình trạng hôn nhân. Có thể thấy, khoản vay của các khách hàng đã có gia đình "Married" vẫn chiếm một tỷ lệ cao trong số các khoản vay bị từ chối. Điều này cho thấy tồn tại một số yếu

tố đặc biệt khi đánh giá khả năng trả nợ của khách hàng đã có gia đình, dẫn đến sự tăng cường trong quyết định từ chối cho vay. Tỷ lệ khoản vay bị từ chối cao tiếp theo là các khoản vay của các khách hàng còn độc thân. Điều này cho thấy sự tác động của tình trạng hôn nhân đến quyết định từ chối cho vay, với tỷ lệ từ chối cao hơn đối với nhóm khách hàng này. Có thể rằng việc không có tình trạng hôn nhân có thể được xem như một yếu tố rủi ro trong việc đánh giá khả năng trả nợ của khách hàng.

Tuy tỷ lệ này không cao như nhóm khách hàng đã có gia đình, nhưng vẫn đáng chú ý vì nó chỉ ra sự ảnh hưởng của tình trạng hôn nhân đến việc vay tiền. Có thể rằng các khách hàng còn độc thân thường có những yếu tố riêng như không có người cùng chia sẻ trách nhiệm tài chính hoặc không có nguồn thu nhập phụ, dẫn đến việc tăng khả năng mắc nợ hay không đủ khả năng trả nợ.

Tỷ lệ từ chối đơn xin vay đối với các khách hàng có tình trạng hôn nhân là Civil marriage có thể có mức độ từ chối trung bình. Có thể do hôn nhân dân sự không được xem là một hình thức hôn nhân truyền thống, khiến một số tổ chức tín dụng hoặc ngân hàng có quan ngại về việc đánh giá khả năng trả nợ của khách hàng.

Tiếp theo, tình trạng hôn nhân là Separated có thể có tỷ lệ từ chối đơn xin vay tương đối cao. Xem xét thấy do việc ly hôn hoặc chia tay thường đi kèm với các tình huống tài chính phức tạp và không chắc chắn, làm tăng nguy cơ khách hàng không có khả năng trả nợ đúng hẹn.

Tình trạng hôn nhân là widow (góa phụ) có thể có tỷ lệ từ chối đơn xin vay thấp hơn so với các tình trạng hôn nhân khác. Điều này có thể bởi vì các góa phụ thường có một nguồn thu nhập ổn định từ việc thừa kế hoặc các nguồn tài chính khác sau khi mất đối tác, làm tăng khả năng trả nợ và đáp ứng các yêu cầu vay.

4.3.4. Kết luận

Home Credit cung cấp các khoản vay(AMT_CREDIT) và đưa ra các đề xuất số tiền trả hàng tháng (AMT_ANNUITY) đa dạng.

Nhiều loại hợp đồng sản phẩm được đưa ra và hợp đồng sản phẩm tiền mặt là cao nhất, tiếp đến là hợp đồng vay tiêu dùng. Đây cũng là hai khoản vay có tỷ lệ trả nợ đúng hạn cao nhất.

Phần lớn khách hàng với khoản thu nhập dưới số tiền 427500 đều có khả năng trả nợ tốt vì số tiền trả góp hàng tháng đều nhỏ hơn số tiền thu nhập, đảm bảo khoản vay có thể trả đúng hạn.

Số tiền cấp vay trung bình cao nhất của Home Credit là cho khách hàng Đã có bằng Đại học và thuộc nhóm có cấp bậc giáo dục Giáo dục đại học. Với các khoản vay thuộc các nhóm khách hàng có cấp bậc giáo dục thấp hơn cũng có số tiền cấp vay thấp hơn. Cho thấy cấp bậc giáo dục là một yếu tố ảnh hưởng đến đánh giá cấp vay của Home Credit.

Mối quan hệ ảnh hưởng giữa tình trạng hôn nhân và việc từ chối cho vay là có tồn tại. Tuy nhiên, riêng với tình trạng hôn nhân là Đã kết hôn, tỷ lệ từ chối cho vay là cao nhất, cho thấy việc đánh giá với các khoản vay thuộc về các khách hàng có tình trạng hôn nhân này còn được đánh giá dựa trên nhiều yếu tố khác, để đảm bảo việc tin cậy khi cho vay với các khách hàng này.

“RATE_INTEREST_PRIMARY” và “RATE_INTEREST_PRIVILEGED” có tỷ lệ giá trị missing cao, mặc dù giá trị chênh lệch không quá lớn, nhưng lại là thuộc tính đặc trưng của các hợp đồng vay tiêu dùng thuộc danh mục POS. Có thể xem xét hai cột này nếu sử dụng phân tích liên quan đến danh mục POS.

Các giá trị ngoại lệ bao gồm: 365243 là giá trị ngoại lai của 5 thuộc tính liên quan đến thông tin ngày DAYS_TERMINATION, DAYS_LAST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_FIRST_DUE, DAYS_FIRST_DRAWING. Tuy nhiên giá trị này vẫn mang ý nghĩa đối với khoản vay vòng (Revolving loans) hay các khoản vay còn hoạt động sẽ có giá trị 365243 ở cột DAYS_TERMINATION.

Các dòng giá trị ngoại lai “XNA” và “XAP” ở các biến phân loại chiếm tỷ lệ phần trăm khá lớn và có thể được coi là một cách thay thế các giá trị missing value khi giá trị thực tế không được biết hoặc không xác định.

TP.HCM, ngày 15 tháng 6 năm 2023

4.4. Bảng installment_payments

4.4.1. Tổng quan về bảng

Bảng “installment_payments” cung cấp thông tin chi tiết về các khoản trả góp của khoản vay trước đó, bao gồm các ngày trả góp, số tiền đã trả,...Bảng “installment_payments” bao gồm 13605401 hàng, 8 cột. Mỗi hàng đại diện cho khoản vay trước với các thông tin khác nhau về phiên bản lịch sử trả góp, khoản trả góp đang quan sát...

Bảng “installment_payments” chỉ chứa các biến số, không có các biến phân loại.

4.4.2. Khám phá các biến

Thuộc tính	**Mô tả**	**Ghi chú**
SK_ID_PREV	ID của khoản tín dụng trước đó	Mỗi đơn vay hiện tại có thể có nhiều đơn vay liên quan trước đó
SK_ID_CURR	ID của khoản vay hiện tại	
NUM_INSTALMENT_VERSION	Phiên bản lịch trả góp của khoản vay trước đó. Giá trị 0 đại diện cho thẻ tín dụng. Thay đổi phiên bản lịch trả góp từ tháng này sang tháng khác cho thấy một số thông số trong lịch trả góp đã thay đổi.	- Giá trị 0: Đại diện cho thẻ tín dụng. Khi "NUM_INSTALMENT_VERSION" có giá trị 0, nghĩa là khoản vay trước đó là một khoản vay liên quan đến thẻ tín dụng. Thẻ tín dụng thường không có lịch trả góp cụ thể, mà người dùng có thể trả góp theo một số tiền tùy ý hoặc toàn bộ số tiền đã sử dụng trên thẻ. - Các giá trị khác của "NUM_INSTALMENT_VERSION" có thể chỉ ra các phiên bản lịch trả góp khác nhau, tương ứng với các cách thức trả góp

		khác nhau cho khoản vay trước đó. Các phiên bản này có thể phản ánh các biến đổi hoặc thay đổi trong lịch trả góp qua thời gian, ví dụ như sự thay đổi về số tiền trả góp, cách thức tính lãi suất, hoặc các điều khoản khác liên quan đến việc trả góp.
NUM_INSTALMENT_NUMBER	Khoản trả góp đang quan sát. Chỉ ra khoản trả góp thứ mấy trong lịch trả góp.	- Một đơn đăng ký vay trước đó sẽ được theo dõi và quan sát trên nhiều khoản trả góp. Do đó, một đơn đăng ký vay trước đó sẽ có nhiều dòng giá trị với các khoản trả góp được quan sát.
DAYS_INSTALMENT	Phần trả góp của khoản tín dụng trước đó dự kiến phải được thanh toán	
DAYS_ENTRY_PAYMENT	Thời gian các khoản tín dụng trước đó được thực sự thanh toán	
AMT_INSTALMENT	Số tiền góp đã được quy định cho khoản vay trước đó trong khoản góp này	
AMT_PAYMENT	Số tiền mà khách hàng đã thực sự thanh toán cho khoản tín dụng trước trong khoản trả góp này	

Bảng 4.10. Bảng khám phá và mô tả các biến trong bảng dữ liệu "instalments_payments"

4.4.2.1. Xem xét và kiểm tra các giá trị rỗng

Column	Total	Percent
DAYS_ENTRY_PAYMENT	2905	0.021352
AMT_PAYMENT	2905	0.021352

Bảng 4.11. Bảng mô tả tỷ lệ missing data theo từng thuộc tính trong bảng "instalments_payments"

Giá trị ở 2 thuộc tính "DAYS_ENTRY_PAYMENT" (Các khoản tín dụng trước đó được thực sự thanh toán) và "AMT_PAYMENT" (Số tiền mà khách hàng đã thực sự thanh toán cho khoản tín dụng trước trong khoản trả góp này) rỗng, lý do có thể xuất phát từ việc người vay đã không thanh toán khoản vay vào kỳ trả góp đang theo dõi. Do đó, không có thông tin về thời gian khoản vay ở kỳ trả góp này được thực sự thanh toán và số tiền mà người vay thực sự thanh toán. Tuy nhiên, tỷ lệ phần trăm giá trị rỗng ở hai thuộc tính này rất nhỏ, có thể loại bỏ được.

4.4.2.2. Phân tích các biến

a. Thống kê mô tả

	AMT_INSTALMENT	AMT_PAYMENT
count	13605401	13602496
mean	17051	17238
std	50570	54736
min	0	0

25%	4226	3398
50%	8884	8126
75%	16710	16108
max	3771488	3771488

Bảng 4.12. Bảng thống kê mô tả các biến trong bảng dữ liệu "instalments_payments"

Thuộc tính "AMT_INSTALMENT"

Thuộc tính "AMT_INSTALMENT" hay khoản trả góp dự kiến mỗi kỳ có giá trị trung bình là 17051, thuộc khoảng tứ phân vị thứ 4, cho thấy giá trị trung bình cao. Có thể có một số khách hàng có khả năng và thu nhập cao, hoặc có các khoản vay lớn.

Giá trị nhỏ nhất của khoản trả góp dự kiến bằng 0. Có thể xảy ra các trường hợp thông tin về số tiền trả góp mỗi kỳ (AMT_INSTALMENT) bị thiếu hoặc không được cung cấp đầy đủ. Trong trường hợp này, giá trị mặc định có thể là 0.

Giá trị lớn nhất của khoản trả góp dự kiến trong kỳ lớn nhất là 3771488. Đây là một khoản trả góp dự kiến cao, có thể khoản vay này thuộc về các khách hàng có nhu cầu vay tài chính với khoản tiền lớn hoặc số kỳ trả góp ít hay thời hạn trả góp ngắn nên số tiền trả góp mỗi kỳ cao.

Độ lệch chuẩn của khoản khoản vay trả góp dự kiến mỗi kỳ cao, cho thấy mức độ chênh lệch trong các khoản trả góp. Điều này có thể chỉ ra rằng có sự đa dạng về mức độ trả góp của khách hàng trong tập dữ liệu.

Thuộc tính "AMT_PAYMENT"

Quan sát thấy, giá trị trung bình khoản trả góp thực sự của khách hàng trên kỳ trả góp được theo dõi lớn hơn giá trị trung bình khoản trả góp theo quy định trên kỳ trả góp được theo dõi. Điều này có thể xuất phát từ lãi suất tích lũy. Khi khoản vay tích lũy lãi suất theo thời gian, số tiền nợ tổng cộng sẽ tăng lên. Để đảm bảo trả đủ số tiền yêu cầu, người vay phải trả thêm để bao gồm cả lãi suất tích lũy này.

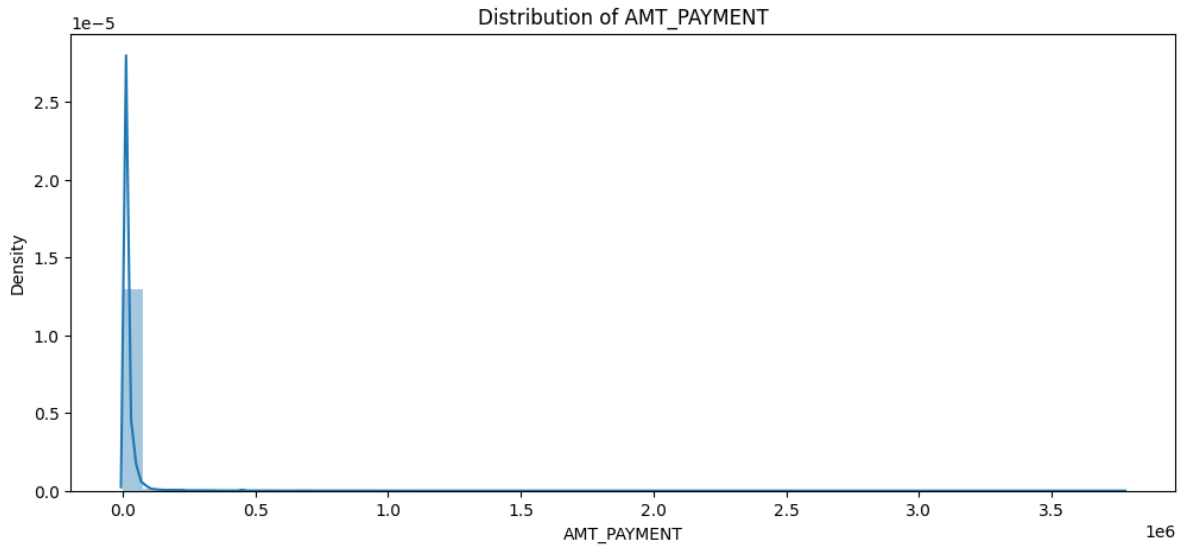
Giá trị nhỏ nhất của khoản trả góp thực sự của kỳ bằng 0 cho thấy có các đơn vay không thực hiện đúng nghĩa vụ thanh toán, thanh toán không đúng với số tiền trả góp

mỗi kỳ theo quy định. Và cũng cần xem xét và đặt câu hỏi về các đơn vay này, có thể đơn vay này sẽ có rủi ro về khả năng thanh toán khoản vay.

Giá trị lớn nhất của khoản trả góp thực sự trong kỳ lớn nhất là 3771488, bằng với giá trị lớn nhất của khoản trả góp dự kiến trong kỳ, cho thấy có tồn tại các đơn vay trả đúng với số tiền trả góp mỗi kỳ theo quy định. Qua đó cũng đánh giá được, với số tiền trả góp lớn như vậy, cũng đánh giá được các đơn vay có giá trị này có khả năng thanh toán cao.

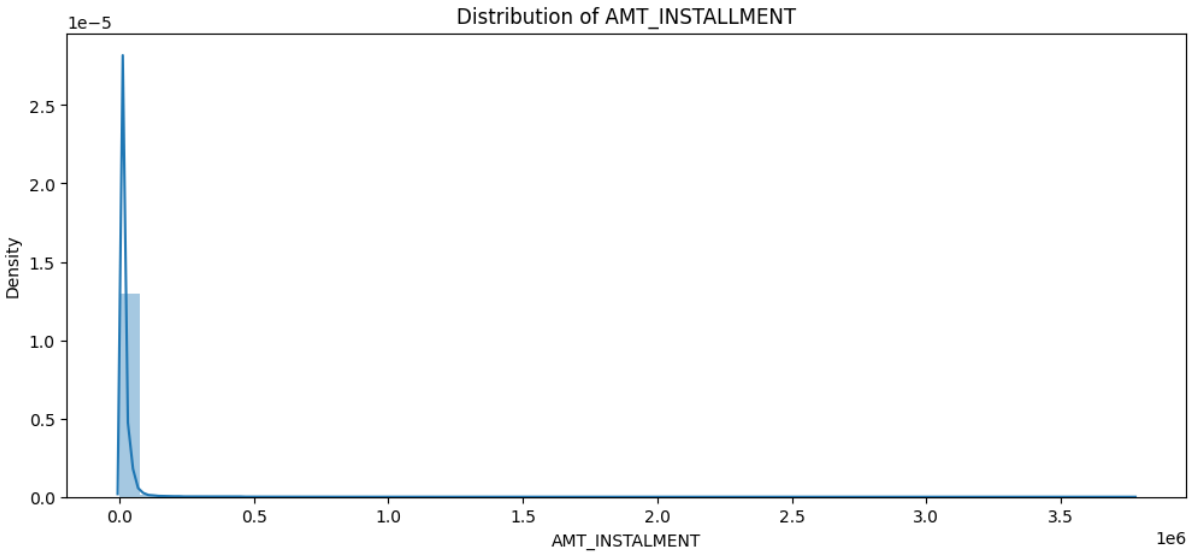
Độ lệch chuẩn lớn hơn giá trị trung bình của số tiền trả góp thật sự mỗi tháng, có thể cho thấy sự biến động mạnh trong số tiền trả góp thực tế. Điều này có thể làm tăng rủi ro về khả năng thanh toán đối với khách hàng hoặc cho thấy sự không ổn định trong việc đảm bảo các khoản trả góp đúng theo dự kiến.

b. Phân phối



Hình 4.28. Biểu đồ phân phối của thuộc tính "AMT_PAYMENT"

(Nguồn: Nhóm tác giả)



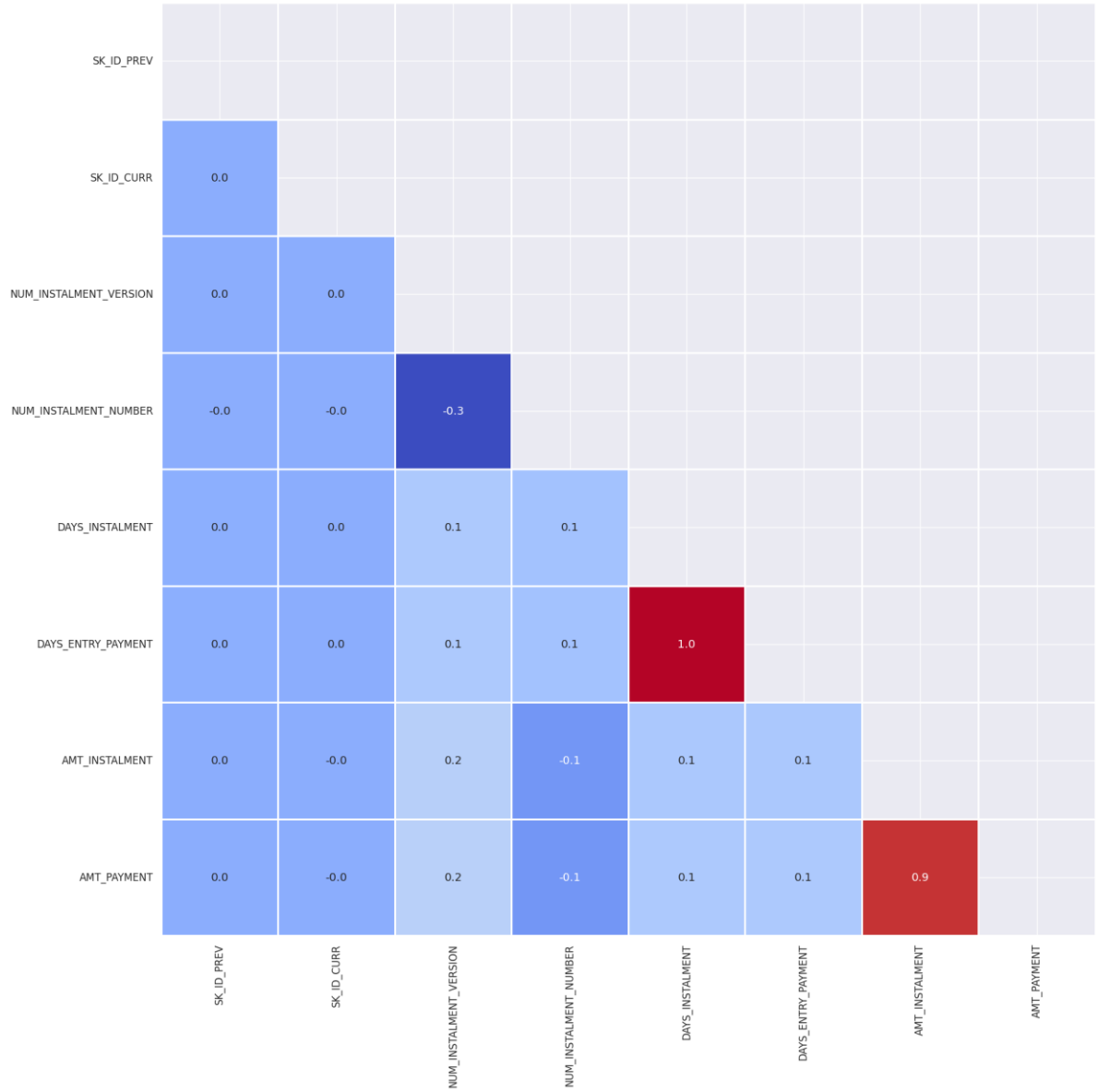
Hình 4.29. Biểu đồ phân phối của thuộc tính "AMT_INSTALMENT"

(Nguồn: Nhóm tác giả)

Hai biểu đồ cho thấy phân phối của "AMT_PAYMENT" (Số tiền mà khách hàng đã thực sự thanh toán cho khoản tín dụng trước trong khoản trả góp này) và "AMOUNT_INSTALLMENT" (Số tiền góp đã được quy định cho khoản vay trước đó trong khoản góp này) có sự biến động và phân tán lớn trong dữ liệu. Tuy nhiên, giá trị của cả hai thuộc tính này tập trung chủ yếu vào khoảng giá trị dưới 100,000. Điều này cho thấy, người vay đăng ký khoản vay nhỏ, do đó số tiền trả góp hàng tháng và cần trả theo quy định không nhiều.

4.4.3. Mối quan hệ với các biến

4.4.3.1 Hệ sống tương quan



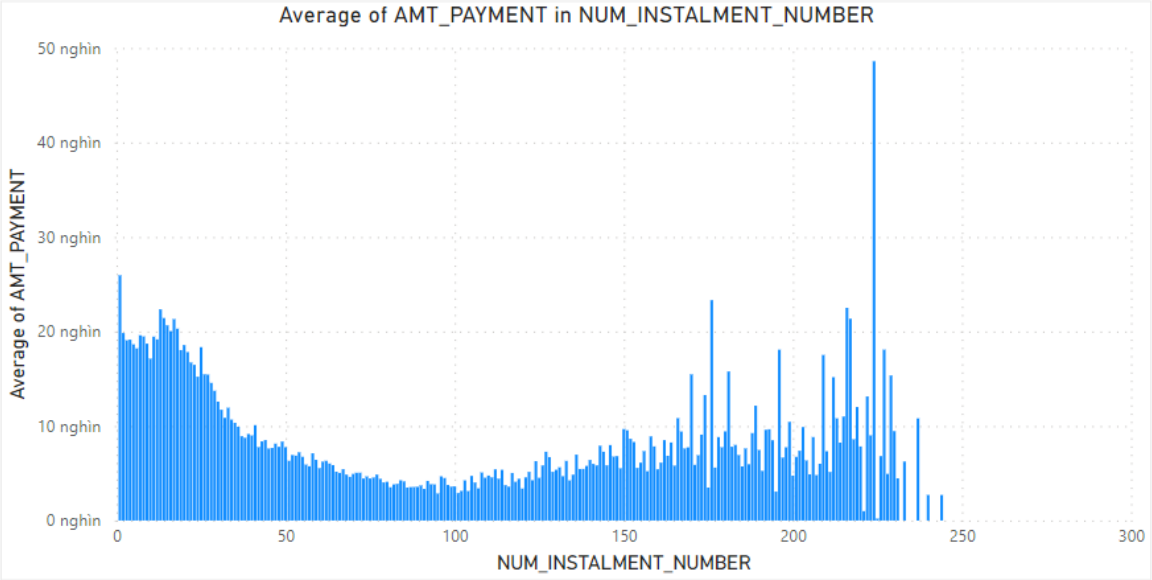
Hình 4.30. Biểu đồ thể hiện tương quan giữa các biến trong bảng "instalments_payments"

(Nguồn: Nhóm tác giả)

Hệ số tương quan giữa thuộc tính "DAYS_ENTRY_PAYMENT" với "DAYS_INSTALMENT" (1.0) và "AMT_PAYMENT" với "AMT_INSTALMENT" (0.9) cao, cho thấy mối tương quan mạnh giữa các biến này. Tuy nhiên, có thể giải thích, giá trị của các biến này hầu như là như nhau, nhưng ý nghĩa của các biến này là

khác nhau. Do đó, không thể loại bỏ một trong hai các cặp thuộc tính có mối quan hệ tương quan mạnh này, để tránh hiện tượng overfitting cho mô hình.

4.4.3.2 Trực quan hóa mối quan hệ giữa các biến



Hình 4.31. Biểu đồ thể hiện giá trị trung bình số tiền trả góp hàng tháng của các kỳ trả góp

Biểu đồ trên cho biết khoản tiền trả góp trung bình người vay thật sự trả qua các kỳ trả góp. Có thể thấy các khoản trả góp ở các kỳ đầu có sự ổn định và chênh lệch không lớn. Điều này có thể cho thấy các khoản trả góp ban đầu được người vay thực hiện theo kế hoạch trả góp và không có những biến động đáng kể. Tuy nhiên, từ kỳ trả góp thứ 25 trở đi, giá trị trung bình của khoản tiền trả góp có sự chênh lệch lớn và giảm mạnh. Lý do có thể xuất phát từ việc người vay có thể gặp khó khăn tài chính khi thời gian trả góp kéo dài, dẫn đến việc trả ít hơn hoặc do lãi suất tích lũy, khoản vay có thể tích lũy lãi suất theo thời gian, dẫn đến tăng số tiền nợ tổng cộng, người vay có thể phải trả nhiều hơn để đảm bảo trả đủ số tiền yêu cầu. Trong giai đoạn gần cuối của quá trình trả góp, giá trị trung bình của khoản tiền trả góp lại có xu hướng tăng dần và tăng mạnh. Có thể người vay cố gắng trả nhiều hơn vào giai đoạn cuối để hoàn thành việc trả góp và tránh việc nợ phát sinh.

4.4.4. Kết luận

Số kỳ càng nhiều, số tiền trả góp hàng tháng càng có khả năng giảm. Tuy nhiên, số tiền trả cho góp cho các kỳ cuối đều có xu hướng tăng, cho thấy người vay vẫn cố gắng hoàn thành việc trả góp và tránh phát sinh thêm nợ.

4.5. Bảng application_train

4.5.1. Tổng quan về bảng

Bảng application_train mô tả thông tin về các ứng dụng vay vốn trong dự án Home Credit. Dưới đây là mô tả tổng quan về bảng application_train:

Bộ dữ liệu application_train có tổng số quan sát là 307,511, tức là bao gồm 307,511 hàng. Mỗi hàng trong bộ dữ liệu application_train tương ứng với một ứng dụng vay vốn của một khách hàng. Mỗi khách hàng gửi một ứng dụng vay vốn và thông tin chi tiết về khách hàng và tình hình tài chính của họ được ghi lại trong hàng đó.

Bộ dữ liệu có tổng số cột là 122, tức là bao gồm 122 biến. Trong số các biến này, có 106 biến số, đại diện cho các đặc trưng có giá trị số như thu nhập, tuổi, số lượng con cái, v.v. Còn lại là 16 biến phân loại, đại diện cho các đặc trưng có giá trị rời rạc như giới tính, tình trạng sở hữu xe, tình trạng sở hữu bất động sản, v.v.

Trong việc phân tích đề tài này, nhóm tập trung vào tìm hiểu các biến có ý nghĩa về mặt nhân khẩu học của khách hàng. Cụ thể, nhóm quan tâm đến 14 biến phân loại như tình trạng hôn nhân, hình thức sở hữu nhà, v.v. và 6 biến số như thu nhập hàng tháng, tuổi, v.v. Những biến này được xem là quan trọng để hiểu và phân tích đặc điểm nhân khẩu học của khách hàng trong quá trình xét duyệt vay vốn.

Với mô tả tổng quan này, nhóm xây dựng một cái nhìn tổng quan về bảng application_train và những đặc trưng quan trọng trong bộ dữ liệu này.

Thuộc tính	Ý nghĩa	Ghi chú
SK_ID_CURR	ID khoản vay hiện tại	

NAME_CONTRACT_TYPE	Loại hợp đồng vay (là vay tiền mặt hay là vay vòng)	- Cash loans (Khoản vay tiền mặt): Đây là loại hợp đồng vay thông thường trong đó khách hàng được cấp một số tiền cụ thể và phải trả lại theo lịch trả góp được định sẵn. - Revolving loans (Credit card): Đại diện cho các khoản vay có tính tái sử dụng. Trong loại hợp đồng này, khách hàng có một khoản tín dụng tối đa đã được phê duyệt và có thể mượn và trả nợ lặp lại trong phạm vi tín dụng đã được xác định. Số tiền trả hàng tháng được tính dựa trên số dư chưa trả và tỷ lệ phần trăm được xác định trước.
CODE_GENDER	Giới tính của khách hàng	- F (Female): Nữ - N (Male): Nam
FLAG_OWN_CAR	Có sở hữu xe hơi hay không?	- Y (Yes):Có - N (No): Không
FLAG_OWN_REALTY	Có sở hữu nhà hay không?	- Y (Yes):Có - N (No): Không
CNT_CHILDREN	Số con của khách hàng	
AMT_INCOME_TOTAL	Tổng thu nhập của khách hàng	
NAME_TYPE_SUITE	Thành viên trong gia đình đi cùng	- "Unaccompanied": Khách hàng nộp đơn một mình, không có ai đi kèm.

TP.HCM, ngày 15 tháng 6 năm 2023

		- "Spouse, partner": Người đi kèm là vợ/chồng hoặc đối tác của khách hàng. - "Family": Người đi kèm là thành viên trong gia đình của khách hàng. - "Children": Người đi kèm là con cái của khách hàng. - "Other_A": Người đi kèm khác (không rõ ràng trong dữ liệu). - "Group of people": Nhóm người đi kèm, có thể là một nhóm người hoặc đại diện một tổ chức. - "Other_B": Người đi kèm khác (không rõ ràng trong dữ liệu, loại khác so với "Other_A").
NAME_INCOME_TYPE	Loại thu nhập của ứng viên	
NAME_EDUCATION_TYPE	Trình độ học vấn của khách hàng	- Secondary / secondary special: Có trình độ học vấn là trung học phổ thông hoặc chuyên ngành trung học phổ thông. - Higher education: Có trình độ học vấn cao hơn, thường là đại học hoặc các trình độ học vấn tương đương. - Incomplete higher: Đã tiến hành học cao học nhưng chưa hoàn thành.

TP.HCM, ngày 15 tháng 6 năm 2023

		- Lower secondary: Có trình độ học vấn là trung học cơ sở. - Academic degree: Đã đạt được một bằng cấp học thuật cao hơn, như tiến sĩ hoặc giáo sư.
NAME_FAMILY_STATUS	Tình trạng hôn nhân	- Single: Giá trị này chỉ ra rằng khách hàng đang độc thân, không kết hôn hoặc không có mối quan hệ hôn nhân hiện tại. - Married: Giá trị này cho thấy khách hàng đã kết hôn và có một mối quan hệ hôn nhân hiện tại. - Civil marriage: Giá trị này chỉ ra rằng khách hàng đang có một mối quan hệ hôn nhân được công nhận pháp lý thông qua hôn nhân dân sự. - Separated: Giá trị này cho thấy khách hàng đã chia tay hoặc ly thân với đối tác hôn nhân hiện tại. - Widow: Giá trị này chỉ ra rằng khách hàng là người góa

		phụ/người đã mất đối tác hôn nhân trước đó. - Unknown: Giá trị này thường xuất hiện khi dữ liệu bị thiếu hoặc không được cung cấp đầy đủ. Điều này có thể xảy ra khi khách hàng không cung cấp thông tin về tình trạng hôn nhân/cuộc sống gia đình hoặc khi dữ liệu không được nhập hoặc ghi lại chính xác.
NAME_HOUSING_TYPE	Loại nhà ở của khách hàng	- "House / apartment": Đại diện cho việc khách hàng sở hữu hoặc thuê một căn hộ hoặc nhà riêng. Đây là loại hình sở hữu nhà phổ biến. - "Rented apartment": Đại diện cho việc khách hàng thuê một căn hộ. - "With parents": Đại diện cho việc khách hàng sống chung với cha mẹ hoặc người nuôi dưỡng. - "Municipal apartment": Đại diện cho việc khách hàng sở hữu hoặc thuê một căn hộ do chính quyền địa phương cung cấp. - "Office apartment": Đại diện cho việc khách hàng sử dụng

		một căn hộ cho mục đích văn phòng. - "Co-op apartment": Đại diện cho việc khách hàng sở hữu hoặc thuê một căn hộ trong một tập thể hợp tác.
DAYS_BIRTH	Số ngày tính từ ngày sinh của ứng viên đến ngày ứng viên nộp đơn vay tiền.	
OCCUPATION_TYPE	Nghề nghiệp của người nộp đơn	
ORGANIZATION_TYPE	Loại tổ chức mà ứng viên đang làm việc	Một số giá trị cần chú ý là: - "Business Entity Type 1": Đại diện cho tổ chức kinh doanh loại 1. Đây có thể là các công ty tư nhân, công ty cổ phần, hoặc các đơn vị kinh doanh khác. - "Business Entity Type 2": Đại diện cho tổ chức kinh doanh loại 2. Đây có thể là các công ty liên doanh, công ty liên kết hoặc các tổ chức kinh doanh có hình thức đặc biệt khác. - "Business Entity Type 3": Đại diện cho tổ chức kinh doanh loại 3. Đây thường là các tổ chức lớn, công ty có quy mô lớn, có thể là các tập đoàn, tập đoàn đa quốc gia hoặc các tổ chức tài chính lớn.

		- "Industry Type 1": Đại diện cho ngành công nghiệp xây dựng. Đây là ngành liên quan đến xây dựng công trình như nhà ở, tòa nhà, cầu đường, vv. - "Industry Type 2": Đại diện cho ngành công nghiệp kinh doanh. Đây là ngành liên quan đến các doanh nghiệp và hoạt động kinh doanh khác nhau. - "Industry Type 3": Đại diện cho ngành công nghiệp thực phẩm. Đây là ngành liên quan đến sản xuất và chế biến thực phẩm. - "Industry Type 4": Đại diện cho ngành công nghiệp năng lượng. Đây là ngành liên quan đến sản xuất, phân phối và sử dụng các nguồn năng lượng như điện, dầu mỏ, khí đốt, vv. - "Industry Type 5": Đại diện cho ngành công nghiệp bảo hiểm. Đây là ngành liên quan đến các hoạt động bảo hiểm và quản lý rủi ro. - "Industry Type 6": Đại diện cho ngành công nghiệp bất động sản. Đây là ngành liên quan đến mua bán, cho thuê và quản lý bất động sản.

TP.HCM, ngày 15 tháng 6 năm 2023

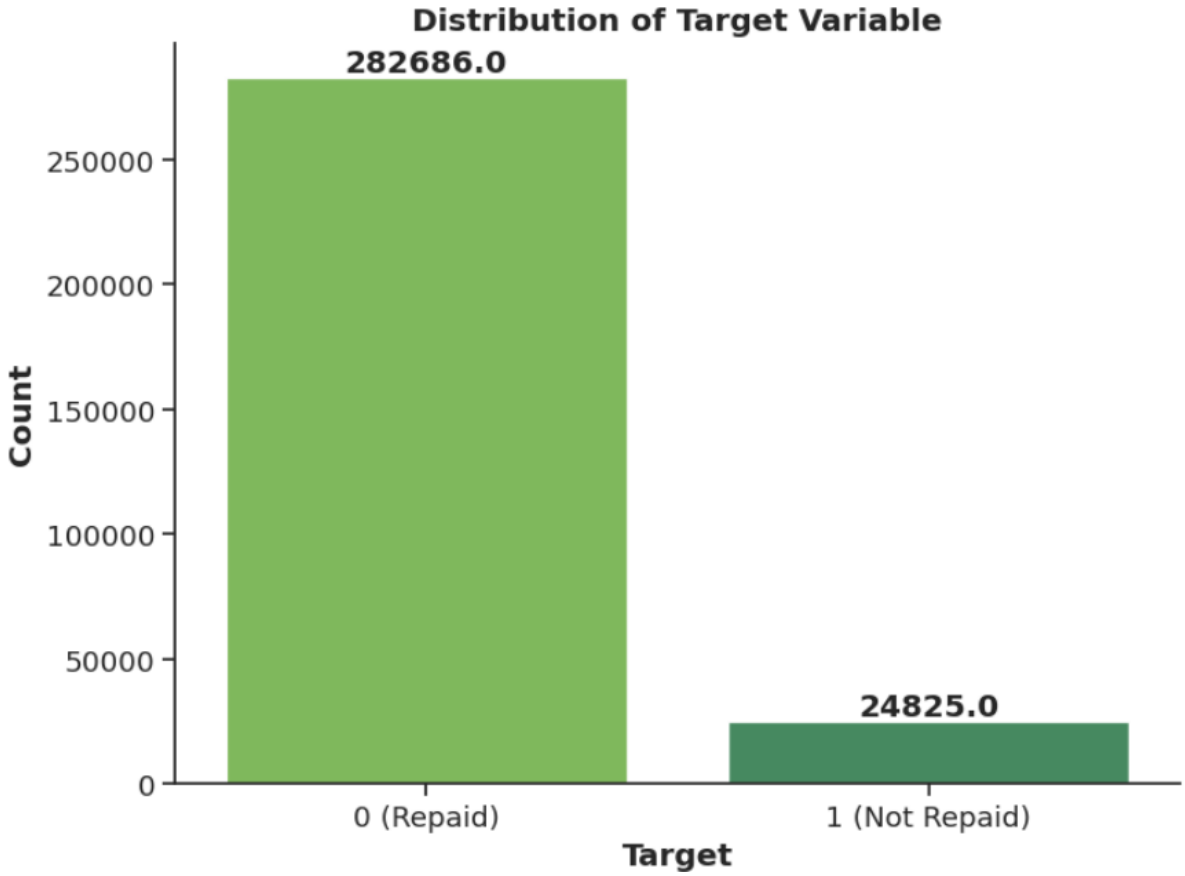
		- "Industry Type 7": Đại diện cho ngành công nghiệp dịch vụ. Đây là ngành liên quan đến cung cấp các dịch vụ khác nhau như dịch vụ vận chuyển, dịch vụ lao động, vv. - "Industry Type 8": Đại diện cho ngành công nghiệp giáo dục. Đây là ngành liên quan đến việc cung cấp dịch vụ giáo dục và đào tạo. - "Industry Type 9": Đại diện cho ngành công nghiệp quân đội. Đây là ngành liên quan đến các hoạt động quân sự và liên quan đến quốc phòng. - "Industry Type 10": Đại diện cho ngành công nghiệp công nghệ thông tin. Đây là ngành liên quan đến công nghệ thông tin, phần mềm, mạng và các dịch vụ liên quan. - "Industry Type 11": Đại diện cho ngành công nghiệp địa ốc. Đây là ngành liên quan đến mua bán, cho thuê - "Industry Type 12": Đại diện cho ngành công nghiệp xã hội. Đây là ngành liên quan đến các tổ chức xã hội và các hoạt động

		liên quan đến phúc lợi xã hội, giúp đỡ cộng đồng. - "Industry Type 13": Đại diện cho ngành công nghiệp khác. Đây là một giá trị chung để đại diện cho các ngành công nghiệp không thuộc các danh mục cụ thể đã được liệt kê trước đó.
FONDKAPREMONT_MODE	Chế độ quản lý quỹ tài sản tương ứng với khu vực sống của người vay.	
WALLSMATERIAL_MODE	Loại vật liệu xây dựng tường trong nhà của người vay.	
HOUSETYPE_MODE	Loại hình nhà ở của người vay.	
CNT_FAM_MEMBERS	Số lượng thành viên trong gia đình của người vay.	
TARGET	Biến mục tiêu (target) có giá trị 0 hoặc 1, biểu thị cho việc khách hàng có thể trả nợ đúng hạn (0) hay không (1).	

Bảng 4.13. Bảng khám phá và mô tả các biến trong bảng dữ liệu "*application_train*"

4.5.2. Khám phá các biến

Phân phối của biến 'Target'



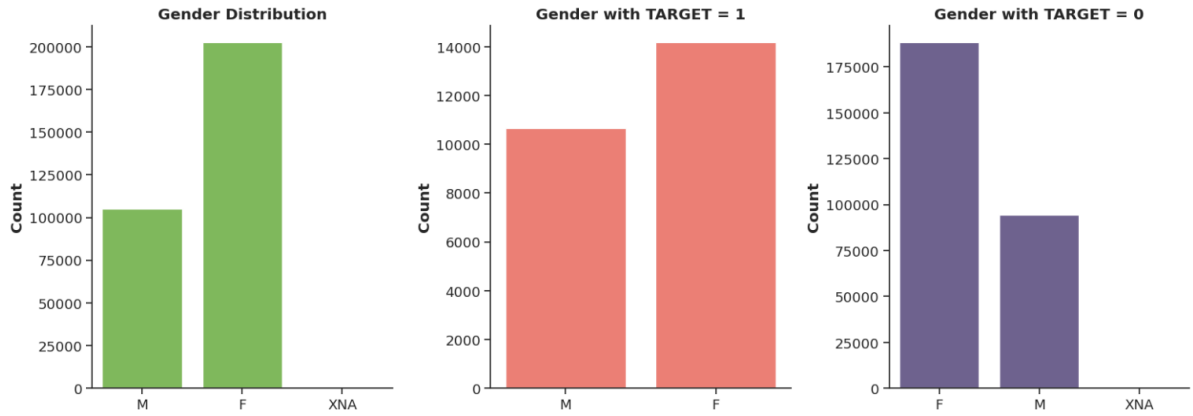
Hình 4.32. Phân phối của biến target trong bảng application_train

(Nguồn: Nhóm tác giả)

Biểu đồ sự phân phối của biến 'TARGET' trong bảng dữ liệu 'application_train' cho thấy số lượng khách hàng có giá trị của biến target là 0 ('TARGET' = 0) là 282,386, và số lượng khách hàng không trả nợ đúng hạn hoặc gặp khó khăn trong việc trả nợ ('TARGET' = 1) là 24,825.

Nhìn chung, sự khác biệt giữa hai giá trị của biến 'TARGET' là khá lớn. Số lượng khách hàng trả nợ không đúng hạn chiếm một tỷ lệ nhỏ so với tổng số khách hàng trong bảng dữ liệu nhưng vẫn tạo ra mức độ rủi ro cho doanh nghiệp.

Giới tính của khách hàng



Hình 4.33. Giới tính của khách hàng

(Nguồn: Nhóm tác giả)

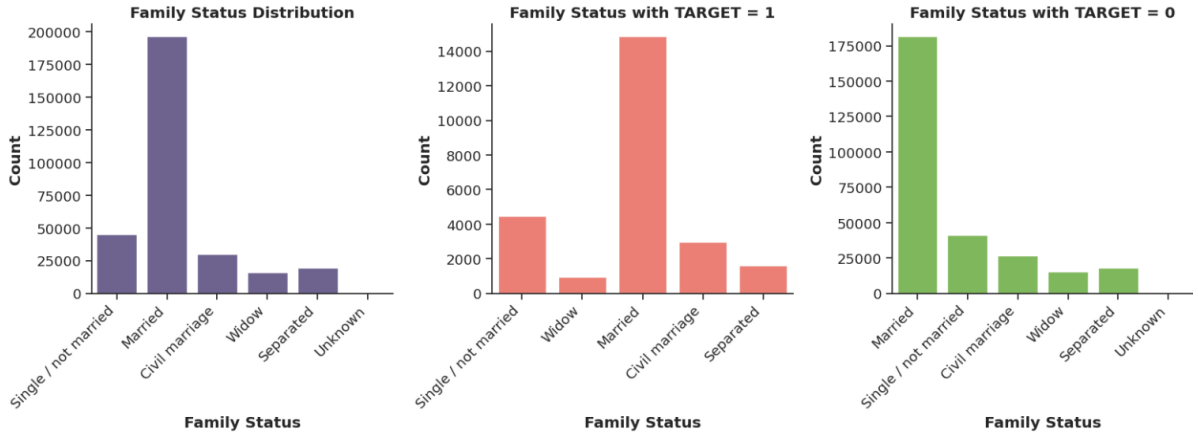
Biểu đồ thể hiện phân phối giới tính của các khách hàng trong bảng dữ liệu "application_train" và biểu diễn phân phối số lượng khách hàng với các giới tính theo từng tình trạng vay của khách hàng.

Bảng dữ liệu "application_train" cho thấy số lượng đơn xin vay của khách hàng nữ chiếm tỷ lệ lớn hơn so với khách hàng nam, thể hiện các khách hàng nữ có xu hướng và nhu cầu vay cao hơn so với các khách hàng nam.

Biểu đồ [2] cho thấy tỷ lệ trường hợp không thể thanh toán khoản vay cao hơn đối với khách hàng nữ so với khách hàng nam. Điều này cho thấy rằng phụ nữ có xu hướng gặp khó khăn hơn trong việc trả nợ và có khả năng rơi vào tình trạng không thể thanh toán cao hơn.

Tuy nhiên, ở biểu đồ [3], khách hàng nữ lại có tỷ lệ trả nợ cao hơn gấp 2 lần so với khách hàng nam. Có thể thấy, mặc dù khách hàng nữ có tỷ lệ không thể thanh toán cao hơn, nhưng khi có khả năng trả nợ, họ có xu hướng trả nợ đầy đủ hơn so với khách hàng nam.

Trạng thái gia đình của khách hàng



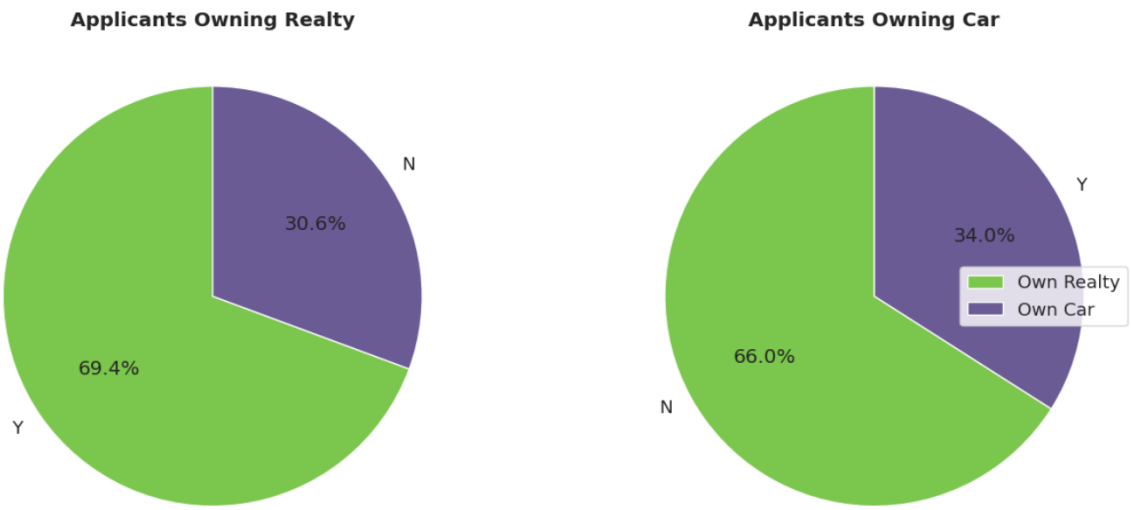
Hình 4.34. Trạng thái hôn nhân của khách hàng

(Nguồn: Nhóm tác giả)

Biểu đồ thể hiện phân phối tình trạng hôn nhân của các khách hàng trong bảng dữ liệu "application_train" và biểu diễn phân phối số lượng khách hàng trong từng loại tình trạng hôn nhân với từng tình trạng vay của khách hàng.

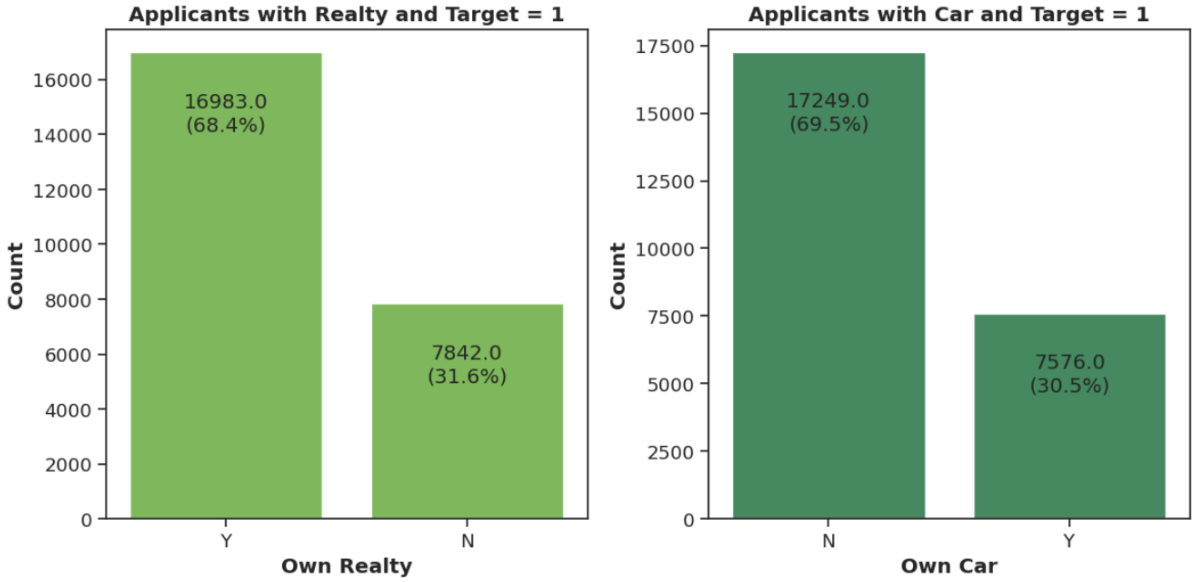
Biểu đồ cho thấy khách hàng sử dụng dịch vụ vay của Home Credit đã có gia đình chiếm một tỷ lệ rất cao trong bảng dữ liệu "application_train". Khách hàng Đã có gia đình vẫn chiếm một tỷ lệ rất cao trong các khoản vay trả nợ và không trả nợ. Các tình trạng hôn nhân khác cũng có tỷ lệ trong các khoản vay trả nợ gần như tỷ lệ trong các khoản vay không trả nợ. Cho thấy tình trạng hôn nhân có ảnh hưởng đáng kể đến khả năng trả nợ của khách hàng. Tuy tỷ lệ trả nợ và không trả nợ của khách hàng đã có gia đình đều cao, nhưng cần chú ý rằng tỷ lệ không trả nợ có thể cao hơn trong nhóm này. Điều này có thể được giải thích bằng việc các khách hàng đã có gia đình có thể đối mặt với nhiều trách nhiệm tài chính khác nhau, như nuôi con cái, trả tiền thuê nhà, trả học phí, v.v. Ngoài ra, các tình trạng hôn nhân khác như civil marriage, separated, và widow cũng có tỷ lệ trả nợ và không trả nợ tương đối gần nhau, cho thấy rằng các yếu tố khác có thể ảnh hưởng đến khả năng trả nợ của khách hàng như tình trạng tài chính cá nhân, lịch sử tín dụng, hay mức độ ổn định trong cuộc sống.

Tỷ lệ khách hàng sở hữu nhà và xe



Hình 4.35. Tỷ lệ khách hàng sở hữu nhà (biểu đồ bên trái) và sở hữu xe (biểu đồ bên phải)

(Nguồn: Nhóm tác giả)



Hình 4.36. Tỷ lệ khách hàng sở hữu nhà không trả nợ đúng hạn (bên trái) và khách hàng sở hữu xe không trả nợ đúng hạn (bên phải)

(Nguồn: Nhóm tác giả)

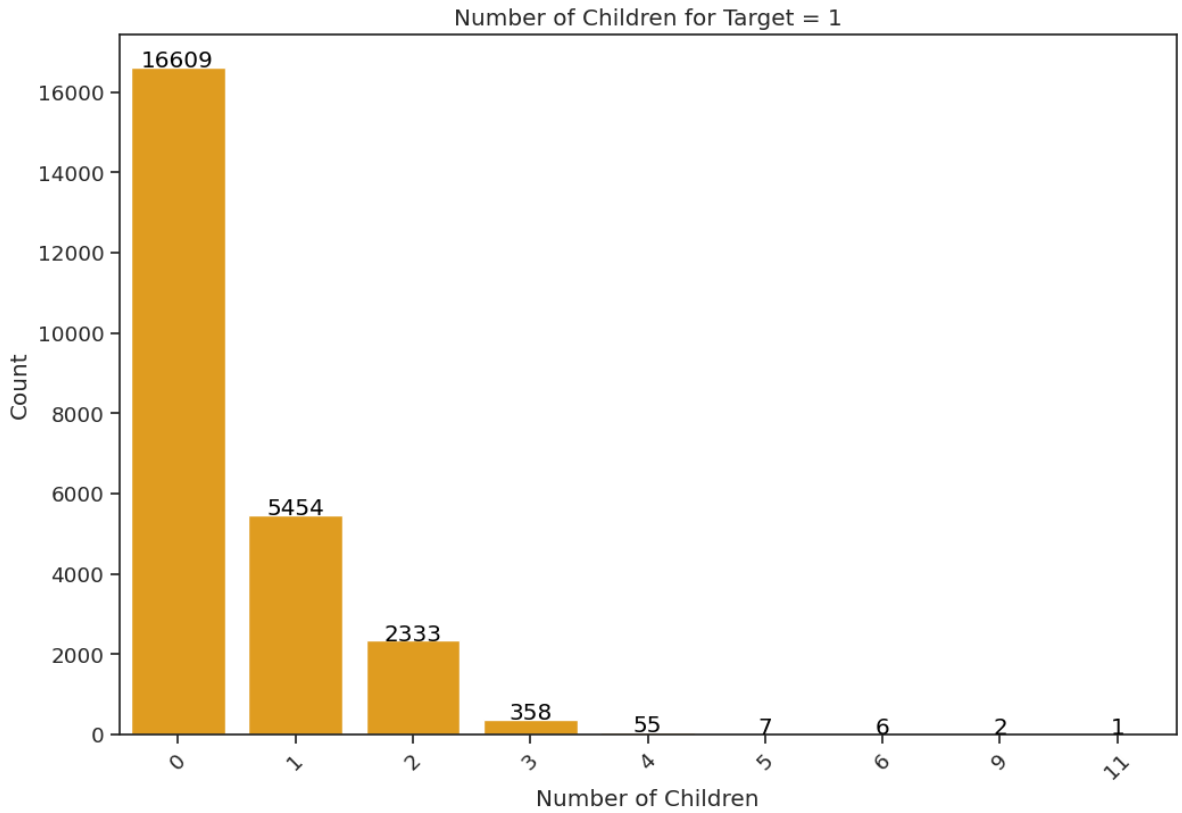
Biểu đồ thể hiện phân phối tình trạng có nhà và có xe của các khách hàng trong bảng dữ liệu "application_train" và biểu diễn phân phối số lượng khách hàng trong từng loại tình trạng có nhà và có xe với từng tình trạng vay của khách hàng.

Có thể thấy, số lượng lớn các khoản vay trong bảng dữ liệu "application_train" là thuộc về các khách hàng không có xe và các khách hàng sở hữu nhà.

Các khoản vay không có khả năng thanh toán lại thuộc về một số lượng lớn các khách hàng có nhà. Một giải thích có thể là các khách hàng sở hữu nhà thường có các khoản vay lớn hơn, liên quan đến việc mua bất động sản hoặc đầu tư vào các dự án. Điều này có thể dẫn đến mức trách nhiệm tài chính cao hơn và áp lực trả nợ lớn hơn cho các khách hàng này. Nếu không có kế hoạch tài chính phù hợp hoặc gặp khó khăn về thu nhập, khách hàng có thể gặp khó khăn trong việc trả các khoản vay này.

Và các không vay không có khả năng thanh toán lại thuộc về một số lượng lớn các khách hàng không xe. Điều này có thể cho thấy một mối liên hệ giữa việc sở hữu xe và khả năng trả nợ của khách hàng. Có thể rằng khách hàng không có xe gặp khó khăn trong việc thu thập thu nhập đủ để đáp ứng các khoản vay và trả nợ một cách đầy đủ.

Số lượng con cái trong gia đình của ứng viên trả nợ không đúng hạn

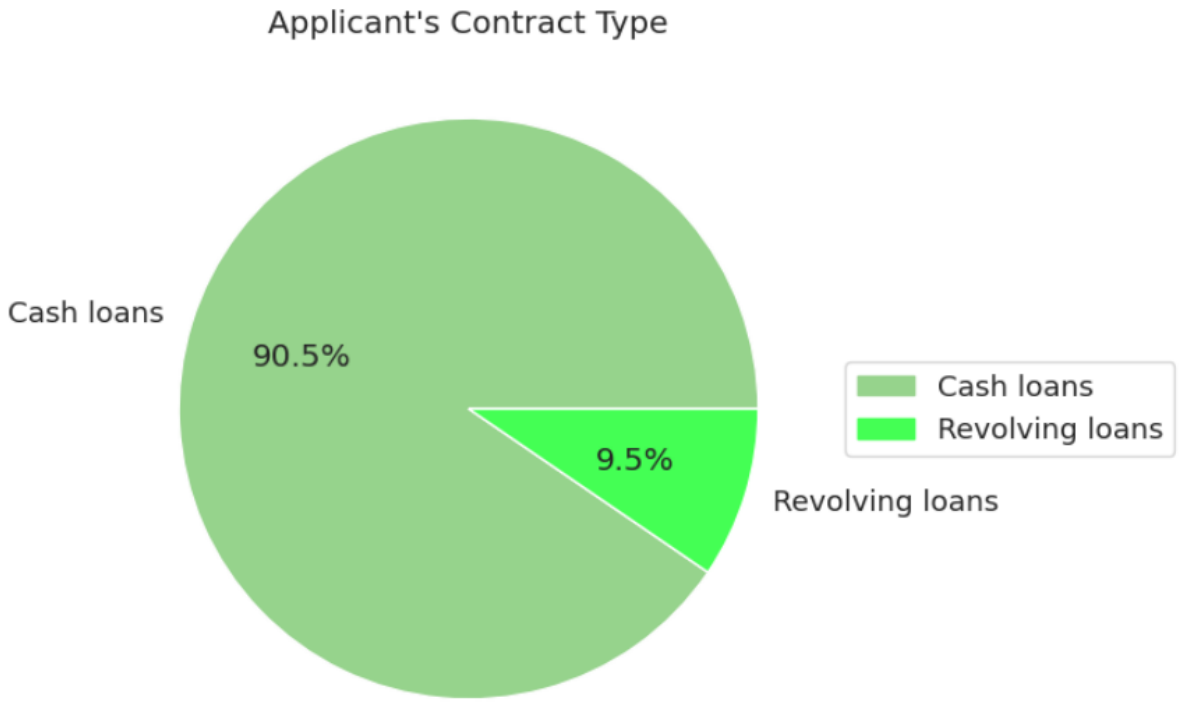


Hình 4.37. Biểu đồ cột thể hiện mối quan hệ giữa số lượng con cái trong gia đình và tình trạng trả nợ không đúng hạn của ứng viên.

(Nguồn: Nhóm tác giả)

Cho thấy các ứng viên không có con có tỷ lệ không trả nợ cao nhất. Lý do có thể là vì vay tiền cho các mục đích cá nhân như mua xe hơi, mua nhà hoặc du lịch. Nếu họ không quản lý tài chính cẩn thận hoặc không đủ khả năng trả nợ, họ có thể rơi vào mức độ nợ cao. Hoặc nếu người chưa có con đang theo học hoặc tiếp tục nâng cao trình độ học vấn, họ có thể phải vay vốn để trang trải chi phí học phí. Số con càng nhiều thì khả năng trả càng cao, xét thấy có số lượng con lớn có thể tạo ra sự ổn định tài chính hơn. Người có gia đình đông đúc có thể có sự phân chia trách nhiệm tài chính, tiết kiệm chi phí thông qua việc mua sắm theo số lượng lớn. Điều này giúp họ có thêm tài chính để trả nợ đúng hạn.

Biểu đồ tỷ lệ hợp đồng của khách hàng

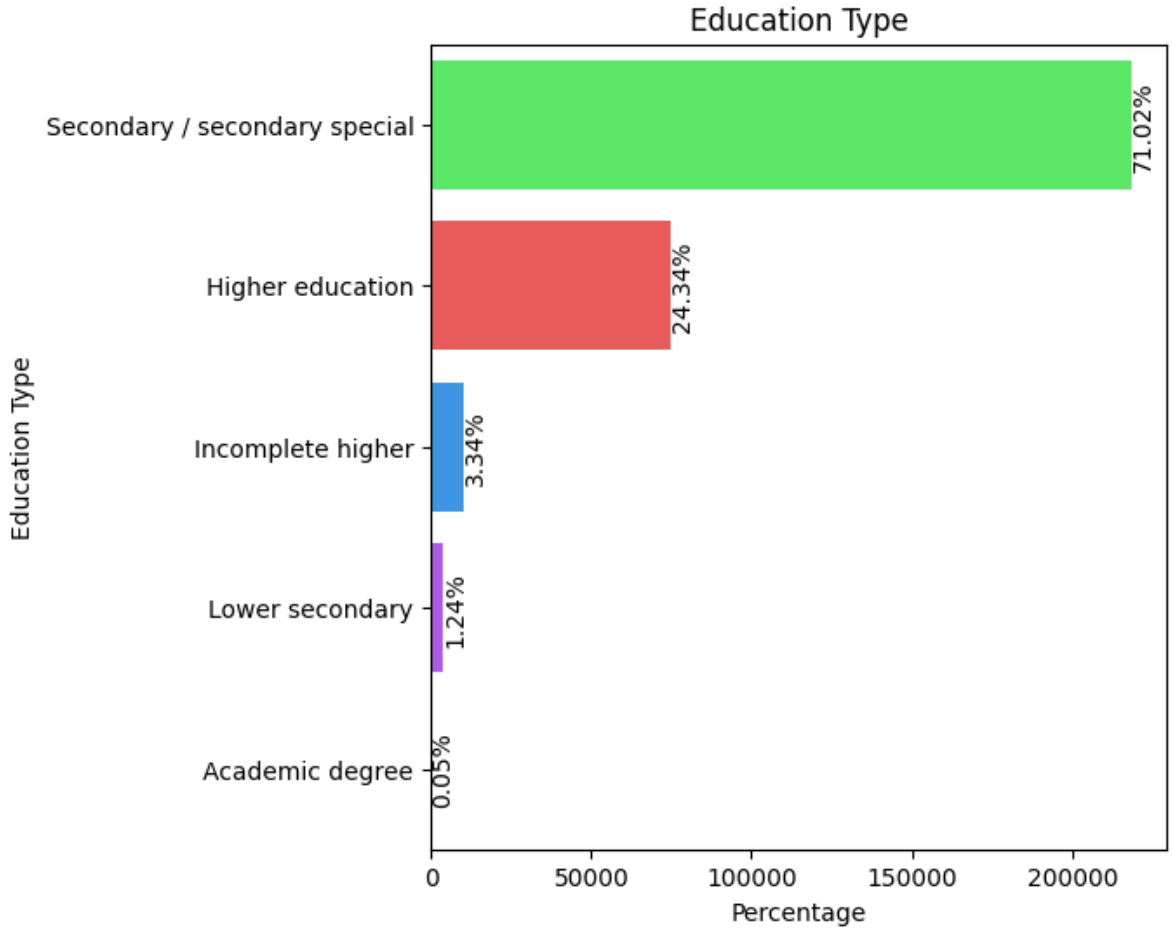


Hình 4.38. Biểu đồ thể hiện tỷ lệ phần trăm các loại hợp đồng trong đơn xin vay của khách hàng.

(Nguồn: Nhóm tác giả)

Có thể thấy, tỷ lệ hợp đồng vay tiền mặt chiếm tỷ lệ lớn, gấp gần 10 lần với hợp đồng vay vòng. Do đó, khách hàng có xu hướng ưa thích và lựa chọn hình thức vay tiền mặt nhiều hơn so với hình thức vay vòng. Lý do có thể là vì một số ưu điểm của khoản vay tiền mặt. Ví dụ như, với cash loan, khách hàng chỉ phải trả nợ theo lịch trình đã được định sẵn và sau khi trả hết số tiền vay, hợp đồng kết thúc, không có rủi ro nợ kéo dài hay việc sử dụng lại số tiền vay như trong revolving loan. Bên cạnh đó, cash loan thường không yêu cầu tình trạng tín dụng tốt như revolving loan. Điều này có nghĩa là người vay có thể dễ dàng đạt được khoản vay tiền mặt mà không cần phải có lịch sử tín dụng hoặc tín dụng tốt. Ngoài ra, trong cash loan, lãi suất thường được cố định trong suốt thời gian vay. Điều này giúp người vay dễ dàng dự tính các khoản chi trả hàng tháng mà không phải lo ngại về sự biến động của lãi suất.

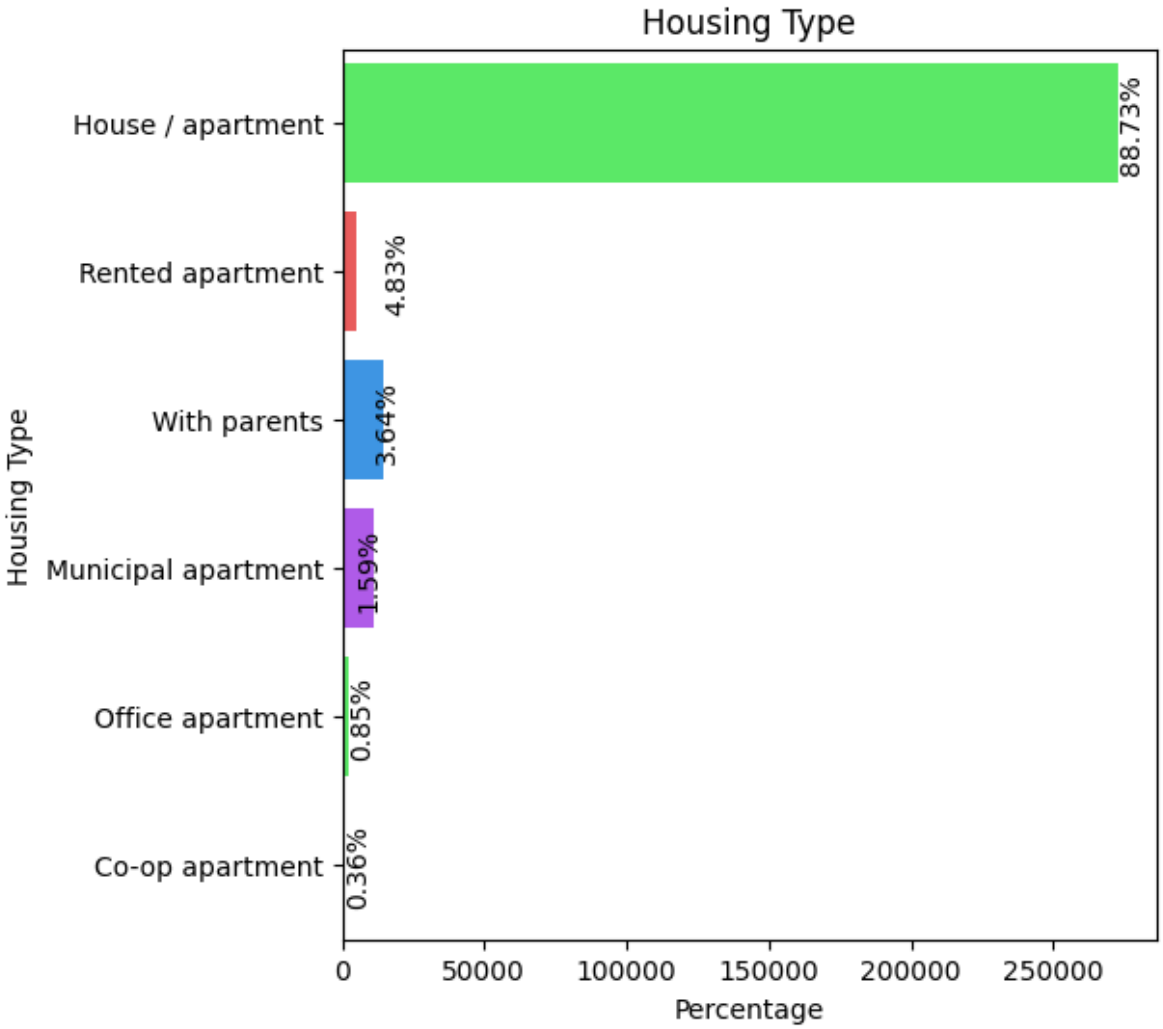
Biểu đồ tỷ lệ các loại hợp đồng của khách hàng và loại nhà của khách hàng



Hình 4.39. Biểu đồ tỷ lệ phần trăm số lượng hợp đồng theo các loại Cấp bậc giáo dục

(Nguồn: Nhóm tác giả)

Biểu đồ trên thể hiện tỷ lệ số lượng các cấp bậc giáo dục của người nộp đơn vay. Có thể thấy, khách hàng có cấp bậc giáo dục "Đã học phổ thông" chiếm tỷ lệ cao nhất, vượt trội so với các cấp bậc giáo dục khác. Những người có cấp bậc "Đã học phổ thông" này có thể đối mặt với nhiều nhu cầu tài chính khác nhau, bao gồm chi trả học phí đại học, chi tiêu cá nhân, chi trả hóa đơn hàng tháng, hoặc đáp ứng nhu cầu tiêu dùng. Điều này có thể giải thích tại sao khách hàng có cấp bậc giáo dục này có tỷ lệ đơn vay cao hơn so với các cấp bậc giáo dục khác. Tiếp theo là khách hàng có cấp bậc giáo dục "Đã học đại học". Thấp hơn là cấp bậc giáo dục "Đang học đại học", "Lower Secondary" và "Đã được cấp bằng học thuật cao hơn". Lý do các đơn vay của khách hàng "Đã được cấp bằng học thuật cao hơn" thấp nhất có thể là vì khách hàng đã đạt được cấp bậc học thuật cao hơn có thể có nhiều cơ hội nghề nghiệp tốt hơn và thu nhập cao hơn, do đó, có thể làm giảm nhu cầu vay tiền hoặc tạo điều kiện thuận lợi để khách hàng trả nợ đúng hạn, dẫn đến tỷ lệ đơn vay thấp hơn trong nhóm này.



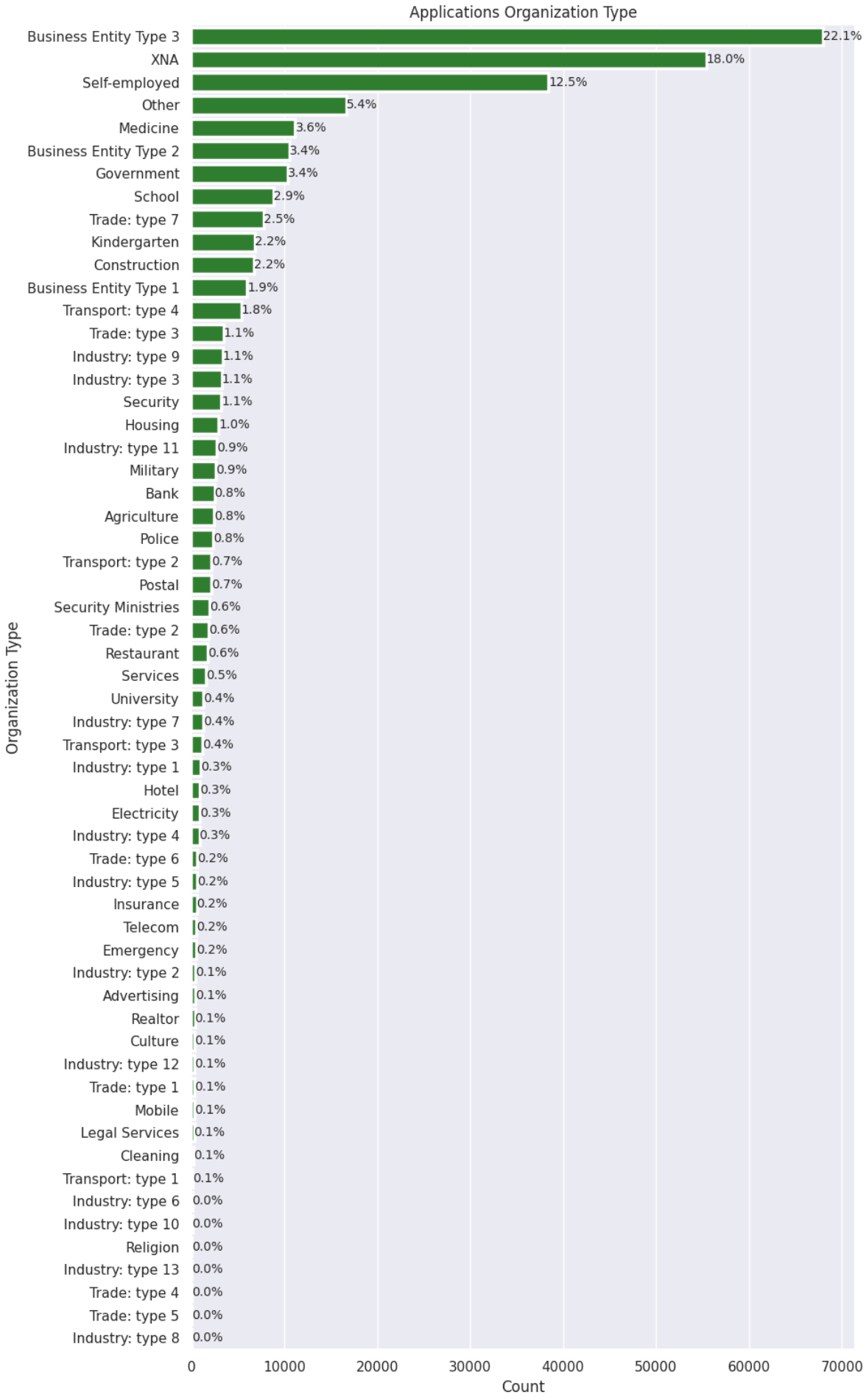
Hình 4.40. Biểu đồ tỷ lệ phần trăm số lượng hợp đồng theo các Loại hình nhà ở của người vay

(Nguồn: Nhóm tác giả)

Biểu đồ dưới đây thể hiện tỷ lệ số lượng khoản vay thuộc về các khách hàng sở hữu loại nhà "Nhà riêng/Căn hộ". Có thể thấy, tỷ lệ này chiếm vị trí hàng đầu và cao hơn đáng kể so với các loại nhà khác. Lý do cho sự ưu thế của khách hàng sở hữu nhà riêng/căn hộ có thể là do một số người có nhu cầu vay tiền để đáp ứng các mục tiêu tài chính cá nhân khác như đầu tư kinh doanh, du lịch, học tập hay thanh toán các khoản nợ khác. Sở hữu nhà/căn hộ có giá trị tài sản đảm bảo giúp họ tiếp cận với khoản vay lớn hơn và có điều kiện vay tốt hơn. Bên cạnh đó, người sở hữu nhà/căn hộ cũng thường có trình độ tài chính tốt hơn so với những người không sở hữu tài sản này. Họ có khả năng đảm bảo trả nợ và được các ngân hàng đánh giá cao về khả năng vay. Do

đó, họ có thể được hưởng ưu đãi lãi suất thấp hơn và điều kiện vay thuận lợi hơn. Ngoài ra, người sở hữu có thể sử dụng nhà/căn hộ như tài sản đảm bảo để đảm nhận rủi ro cho ngân hàng, và do đó, có khả năng vay một số tiền lớn hơn so với những người không có tài sản đảm bảo.

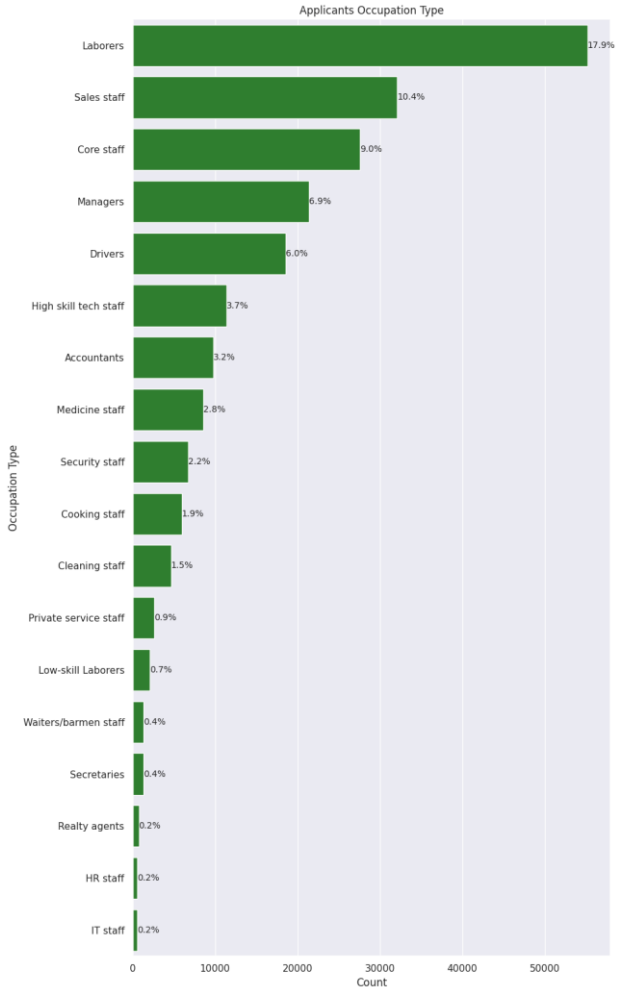
TP.HCM, ngày 15 tháng 6 năm 2023


Applications Organization Type
Business Entity Type 3 22.1%
XNA 18.0%
Self-employed 12.5%
Other 5.4%
Medicine 3.6%
Business Entity Type 2 3.4%
Government 3.4%
School 2.9%
Trade: type 7 2.5%
Kindergarten 2.2%
Construction 2.2%
Business Entity Type 1 1.9%
Transport: type 4 1.8%
Trade: type 3 1.1%
Industry: type 9 1.1%
Industry: type 3 1.1%
Security 1.1%
Housing 1.0%
Industry: type 11 0.9%
Military 0.9%
Bank 0.8%
Agriculture 0.8%
Police 0.8%
Transport: type 2 0.7%
Postal 0.7%
Security Ministries 0.6%
Trade: type 2 0.6%
Restaurant 0.6%
Services 0.5%
University 0.4%
Industry: type 7 0.4%
Transport: type 3 0.4%
Industry: type 1 0.3%
Hotel 0.3%
Electricity 0.3%
Industry: type 4 0.3%
Trade: type 6 0.2%
Industry: type 5 0.2%
Insurance 0.2%
Telecom 0.2%
Emergency 0.2%
Industry: type 2 0.1%
Advertising 0.1%
Realtor 0.1%
Culture 0.1%
Industry: type 12 0.1%
Trade: type 1 0.1%
Mobile 0.1%
Legal Services 0.1%
Cleaning 0.1%
Transport: type 1 0.1%
Industry: type 6 0.0%
Industry: type 10 0.0%
Religion 0.0%
Industry: type 13 0.0%
Trade: type 4 0.0%
Trade: type 5 0.0%
Industry: type 8 0.0%
Organization Type
0
10000
20000
30000
40000
50000
60000
70000
Count

Hình 4.41. Biểu đồ biểu diễn tỷ lệ phần trăm số lượng đơn vay phân phối trong từng loại tổ chức mà ứng viên đang làm việc.

(Nguồn: Nhóm tác giả)

Có thể thấy các đơn vay của các khách hàng làm việc trong tổ chức "Business Entity Type 3" chiếm phần trăm lớn nhất, là các tổ chức lớn, công ty có quy mô lớn, có thể là các tập đoàn, tập đoàn đa quốc gia hoặc các tổ chức tài chính lớn. Các tổ chức lớn thường có nhu cầu vay vốn lớn để đầu tư vào các dự án phát triển, mở rộng kinh doanh, mua sắm thiết bị, nâng cấp hệ thống, hoặc thực hiện các giao dịch mua bán quy mô lớn. Bên cạnh đó, ngân hàng và các tổ chức tài chính thường đánh giá rủi ro khi cấp vay. Các tổ chức lớn thường có lịch sử tín dụng tốt, các quy trình quản lý nội bộ chặt chẽ, và khả năng hoạt động bền vững. Điều này làm cho các đơn vay từ các tổ chức lớn trở nên an toàn hơn và được đánh giá rủi ro thấp hơn, dẫn đến việc tăng tỷ lệ chấp nhận đơn vay. Các đơn vay của các khách hàng làm việc trong tổ chức không có giá trị xác định "XNA" chiếm phần trăm lớn thứ hai. Do đó, không thể phân tích dựa trên giá trị này. Chiếm tỷ lệ phần trăm lớn thứ ba là các đơn vay của khách hàng nghề tự do. Do tính chất của công việc, khách hàng nghề tự do thường có nhu cầu vay vốn để đầu tư vào việc phát triển kinh doanh cá nhân, mở rộng mạng lưới khách hàng hoặc mua sắm thiết bị và công cụ làm việc.



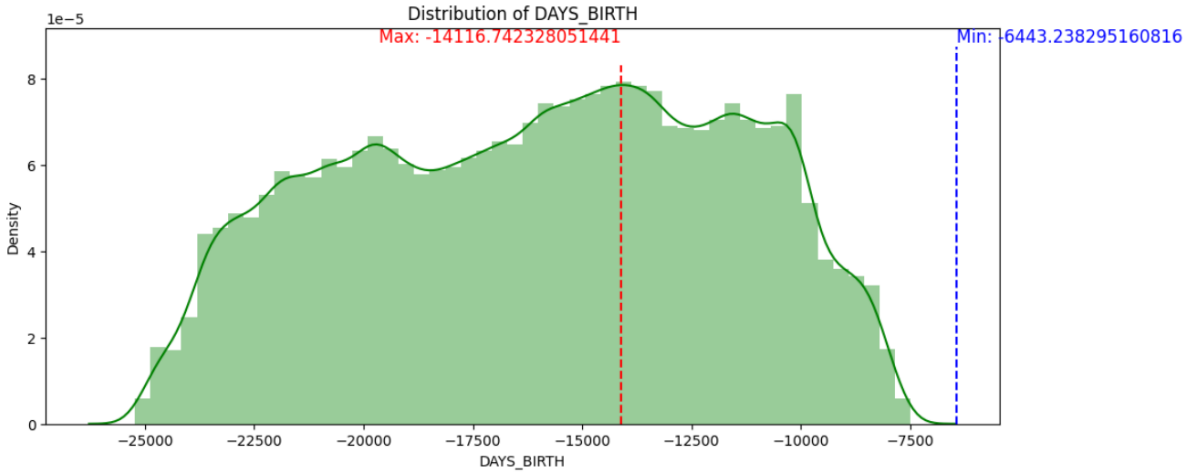
Hình 4.42. Biểu đồ biểu diễn tỷ lệ phần trăm số lượng đơn vay phân phối trong từng nghề nghiệp của người nộp đơn.

(Nguồn: Nhóm tác giả)

Đơn vay thuộc các ngành nghề chiếm tỷ lệ cao nhất lần lượt là Labors, Sales staff, Core staff. Tỷ lệ cao nhất cho đơn vay thuộc ngành nghề "Labors" có thể chỉ ra sự phổ biến của các công việc lao động chủ yếu trong nền kinh tế. Các công việc nhân công có thể bao gồm lao động xây dựng, vận chuyển, công nhân sản xuất, ngành nghề dịch vụ và các công việc thủ công. Những ngành nghề này thường có nhu cầu vay vốn để mua sắm thiết bị, công cụ làm việc hoặc đáp ứng nhu cầu tài chính cá nhân. Tiếp theo là đơn vay thuộc ngành nghề "Sales staff" có thể chỉ ra sự tăng trưởng của ngành bán hàng và hoạt động thương mại. Các ngành nghề bán hàng có thể bao gồm nhân viên bán hàng, đại diện kinh doanh, nhà phân phối, hoặc các chủ cửa hàng. Người làm việc trong ngành này có thể cần vay vốn để mở rộng kinh doanh, tăng cường quảng

cáo, mua hàng tồn kho, hoặc phát triển mạng lưới khách hàng. Chiếm tỷ lệ cao thứ ba cho đơn vay thuộc ngành nghề "Core staff" có thể chỉ ra sự quan trọng của các vị trí nhân viên cốt lõi trong một tổ chức hoặc công ty. Nhân viên cốt lõi có thể là các nhân viên quản lý, chuyên gia trong lĩnh vực chính của công ty hoặc các vị trí quản lý chủ chốt. Đối với các vị trí như vậy, việc vay vốn có thể được sử dụng để đầu tư vào sự phát triển cá nhân, hoặc là là người đại diện vay vốn cho việc đào tạo nâng cao năng lực, mở rộng dự án hoặc thực hiện các giao dịch kinh doanh quan trọng của công ty.

Phân phối của số ngày tính từ ngày sinh của ứng viên đến ngày ứng viên nộp đơn vay tiền



Hình 4.43. Phân phối của số ngày tính từ ngày sinh của ứng viên đến ngày ứng viên nộp đơn vay tiền

(Nguồn: Nhóm tác giả)

Biểu đồ phân phối cho thấy sự phân bố của số ngày tính từ ngày sinh của ứng viên đến ngày ứng viên nộp đơn vay tiền. Trục x của biểu đồ trải dài từ -25000 đến -7500, đại diện cho khoảng tuổi từ 7500 ngày (khoảng 20 tuổi) đến 25000 ngày (khoảng 68 tuổi).

Đường trend của biểu đồ phân phối cho thấy cao nhất là khoảng -14117 (tương đương với tuổi trung bình là 38.7 tuổi) và thấp nhất là khoảng -6443 (tương đương với tuổi trung bình là 17.7 tuổi). Điều này có nghĩa là ứng viên 38.7 tuổi, nhóm tuổi trung niên trong tập dữ liệu chiếm số lượng lớn nhất. Nguyên nhân có thể do các ứng viên trong nhóm này có thu nhập, khoản vay và trách nhiệm tài chính lớn hơn so với các

nhóm tuổi khác.Và ứng viên có tuổi từ 17.7 tuổi chiếm số lượng nhỏ nhất. Điều này có thể do nhóm này chưa tích lũy đủ kinh nghiệm, thu nhập và tài sản để đảm bảo khả năng trả nợ.

4.5.3. Kết luận

- Số lượng khách hàng trả nợ không đúng hạn chiếm một tỷ lệ nhỏ so với tổng số khách hàng trong bảng dữ liệu
- Phụ nữ có xu hướng gặp khó khăn hơn trong việc trả nợ và có khả năng rơi vào tình trạng không thể thanh toán cao hơn. Mặc dù khách hàng nữ có tỷ lệ không thể thanh toán cao hơn, nhưng khi có khả năng trả nợ, họ có xu hướng trả nợ đầy đủ hơn so với khách hàng nam.
- Khách hàng Đã có gia đình vẫn chiếm một tỷ lệ rất cao trong các khoản vay trả nợ và không trả nợ. Điều này cho thấy tình trạng hôn nhân có ảnh hưởng đáng kể đến khả năng trả nợ của khách hàng.
- Các khoản vay không có khả năng thanh toán lại thuộc về một số lượng lớn các khách hàng có nhà. Nếu không có kế hoạch tài chính phù hợp hoặc gặp khó khăn về thu nhập, khách hàng có thể gặp khó khăn trong việc trả các khoản vay này.
- Khách hàng không có xe gặp khó khăn trong việc thu thập thu nhập đủ để đáp ứng các khoản vay và trả nợ một cách đầy đủ.
- Ứng viên không có con có tỷ lệ không trả nợ cao nhất. Số con càng nhiều thì khả năng trả càng cao, xét thấy có số lượng con lớn có thể tạo ra sự ổn định tài chính hơn.
- Khách hàng có xu hướng ưa thích và lựa chọn hình thức vay tiền mặt nhiều hơn so với hình thức vay vòng.
- Khách hàng có cấp bậc giáo dục "Đã học phổ thông" chiếm tỷ lệ cao nhất, vượt trội so với các cấp bậc giáo dục khác.
- Người sở hữu có thể sử dụng nhà/căn hộ như tài sản đảm bảo để đảm nhận rủi ro cho ngân hàng, có khả năng vay một số tiền lớn hơn so với những người không có tài sản đảm bảo.

- Các tổ chức lớn thường có nhu cầu vay vốn lớn để đầu tư vào các dự án phát triển, mở rộng kinh doanh, mua sắm thiết bị, nâng cấp hệ thống, hoặc thực hiện các giao dịch mua bán quy mô lớn.
- Tỷ lệ ứng viên nhóm tuổi trung niên có thu nhập, khoản vay và trách nhiệm tài chính lớn hơn so với các nhóm tuổi khác, chiếm số lượng lớn nhất trong tập dữ liệu. Và ứng viên có tuổi từ 17.7 tuổi, chưa tích lũy đủ kinh nghiệm, thu nhập và tài sản để đảm bảo khả năng trả nợ, chiếm số lượng nhỏ nhất.

TP.HCM, ngày 15 tháng 6 năm 2023

CHƯƠNG 5: THỰC NGHIỆM

5.1. Tiền xử lý dữ liệu

5.1.1. Feature Selections

Với dataset có nhiều bảng và nhiều trường dữ liệu, việc loại bỏ các trường dữ liệu không cần thiết giúp cho quá trình xử lý dữ liệu trở nên hiệu quả hơn và giảm thiểu thời gian tính toán. Việc giảm số lượng trường dữ liệu không cần thiết còn giúp cho việc phân tích dữ liệu dễ dàng hơn trong việc tập trung vào các trường quan trọng và đưa ra những phân tích chính xác và đáng tin cậy hơn. Ngoài ra, đó cũng là cách để giảm thiểu các lỗi dữ liệu và nâng cao độ chính xác của kết quả phân tích.

Dữ liệu chuẩn sẽ bắt đầu từ tập được gộp lại từ application_train và application_test. Đối với tập dữ liệu này chỉ cần lấy mã id khoản vay và thông tin nhân khẩu học của khách hàng bao gồm 21 thuộc tính bên dưới:

- ID của khoản vay hiện tại: 'SK_ID_CURR'
- Các thông tin nhân khẩu học của khách hàng: 'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL' ,'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS' ,'NAME_HOUSING_TYPE', 'DAYS_BIRTH' ,'OCCUPATION_TYPE' ,'ORGANIZATION_TYPE', 'FONDKAPREMONT_MODE', 'WALLSMATERIAL_MODE' ,'HOUSETYPE_MODE', 'CNT_FAM_MEMBERS'

SK_ID_CURR	is_test	is_train	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_TYPE_SUITE	...	NAME_EDUCATION_TYPE
100002	0	1	Cash loans	M	N	Y	0	202500.0	Unaccompanied	...	Secondary / secondary special
100003	0	1	Cash loans	F	N	N	0	270000.0	Family	...	Higher education
100004	0	1	Revolving loans	M	Y	Y	0	67500.0	Unaccompanied	...	Secondary / secondary special

Hình 5.1. Một phần của dữ liệu ban đầu

(Nguồn: Nhóm tác giả)

Sau đó, tập dữ liệu sẽ được merge lại từ bốn bảng còn lại. Điểm chung của 4 bảng này chỉ thể hiện dữ liệu hành vi giao dịch nên khi dữ liệu trong bảng được group lại theo ID khoản vay để merge với dữ liệu ban đầu sẽ lấy theo giá trị trung bình.

- Bảng Previous Applications, ứng với từng SK_ID_CURR đếm tổng số lượng SK_ID_PREV sẽ biết được ID của khoản vay trong hiện tại có bao ID của khoản vay trước ở Home Credit. Sau đó, dữ liệu sẽ được group theo SK_ID_CURR lấy trung bình các đặc trưng và merge với dữ liệu gốc theo SK_ID_CURR. Để dễ phân biệt các đặc trưng nằm ở bảng nào sẽ lấy tên theo kiểu 'prev_' + với tên thuộc tính gốc.
- Làm tương tự với các bảng Previous Installments lấy tên theo kiểu 'i_', bảng Pos Cash Balance lấy tên theo kiểu 'pcb_', bảng Credit Card Balance lấy tên theo kiểu 'cc_bal_'.

Dữ liệu cuối cùng bao gồm 356255 giao dịch với 75 đặc trưng bao gồm nhân khẩu học khách hàng, và hành vi giao dịch của khách hàng.

SK_ID_CURR	is_test	is_train	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_TYPE_SUITE	...	cc_bal_AMT_RECEIVABLE_PRINCIPAL
100002	0	1	Cash loans	M	N	Y	0	>100k	Unaccompanied	...	NaN
100003	0	1	Cash loans	F	N	N	0	>100k	Family	...	NaN
100004	0	1	Revolving loans	M	Y	Y	0	50k-100k	Unaccompanied	...	NaN

Hình 5.2. Một phần của dữ liệu cuối cùng

(Nguồn: Nhóm tác giả)

5.1.2. Làm sạch dữ liệu

Xử lý dữ liệu thiếu (Missing Data Imputation): Dữ liệu thiếu nằm trong các cột chứa thông tin về hành vi và đại diện cho các ô dữ liệu không có giao dịch tương ứng. Để giải quyết vấn đề này, chúng ta sẽ thay thế dữ liệu thiếu bằng giá trị 0.

Trong tập dữ liệu, không có sự xuất hiện của các bản ghi trùng lặp, tức là mỗi dòng trong tập dữ liệu đại diện cho một giao dịch duy nhất

5.1.3. Mã hóa dữ liệu

Đối với dữ liệu phân loại:

Khi làm việc với các biến phân loại trong dữ liệu, việc chuyển đổi chúng thành dạng số để có thể sử dụng trong các thuật toán học máy là rất cần thiết. Hai phương

pháp mã hóa phổ biến là One-hot encoding và Label encoding, và chúng được sử dụng trong các tình huống khác nhau.

- One-hot encoding được sử dụng khi các giá trị của biến phân loại không có thứ tự hay mức độ quan hệ đặc biệt giữa chúng. Với phương pháp này, mỗi giá trị của biến phân loại sẽ được chuyển đổi thành một cột mới trong tập dữ liệu, và các cột này sẽ mang giá trị 0 hoặc 1 tương ứng với sự xuất hiện hay không xuất hiện của giá trị đó trong mẫu dữ liệu. One-hot encoding thích hợp khi các giá trị phân loại là độc lập và không có mối tương quan với nhau.

  Gồm các đặc trưng: 'NAME_EDUCATION_TYPE', 'DAYS_BIRTH' ,'NAME_HOUSING_TYPE','FONDKAPREMONT_MODE','CODE_GEND ER','OCCUPATION_TYPE','WALLSMATERIAL_MODE','NAME_INCOM E_TYPE','HOUSETYPE_MODE','NAME_TYPE_SUITE','NAME_FAMILY _STATUS','ORGANIZATION_TYPE','AMT_INCOME_TOTAL'

- Label encoding được sử dụng khi các giá trị của biến phân loại có một thứ tự hay mức độ quan hệ nhất định. Với phương pháp này, mỗi giá trị của biến phân loại sẽ được gán một số nguyên tương ứng, từ 0 đến (số lượng giá trị - 1). Label encoding thích hợp khi có một sự tương quan thứ tự giữa các giá trị phân loại.

  Gồm các đặc trưng: 'NAME_CONTRACT_TYPE', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY'

Các dữ liệu phân loại trong dữ liệu chính là các dữ liệu nhân khẩu học của khách hàng và các dữ liệu này đều là dữ liệu định danh không có tuần tự hay thứ tự vì vậy đối với đặc trưng có từ 3 giá trị duy nhất khác nhau thì được sử dụng mã hóa One-hot. Còn lại với các đặc trưng có 2 giá trị duy nhất khác nhau thì sử dụng mã hóa Label.

Đối với dữ liệu số:

Khi làm việc với dữ liệu số, việc sử dụng các phép biến đổi như log, sqrt (căn bậc hai) hoặc cbrt (căn bậc ba) để biến đổi dữ liệu ban đầu và đạt được phân phối gần

phân phối chuẩn hơn. Điều này rất hữu ích khi áp dụng các phương pháp thống kê hoặc các thuật toán học máy dựa trên giả định về phân phối chuẩn.

Skewness (độ lệch) là một độ đo thống kê để đánh giá sự lệch của phân phối dữ liệu. Nếu dữ liệu có skewness lớn hơn 0, tức là có lệch dương, thì có xu hướng có giá trị cao hơn trung vị và đuôi dài hơn về phía phải. Ngược lại, nếu dữ liệu có skewness nhỏ hơn 0, tức là có lệch âm, thì có xu hướng có giá trị thấp hơn trung vị và đuôi dài hơn về phía trái.

StandardScaler là một kỹ thuật chuẩn hóa dữ liệu số bằng cách loại bỏ trung bình và chia độ lệch chuẩn của các biến. Bằng cách sử dụng StandardScaler, chúng ta có thể đảm bảo rằng các biến có cùng đơn vị đo lường và phạm vi giá trị tương tự nhau. Điều này có lợi cho việc so sánh và áp dụng các thuật toán học máy phụ thuộc vào các giá trị tương đối của các biến.

Vì vậy, quy trình tổng quát khi làm việc với dữ liệu số là:

- Kiểm tra skewness của dữ liệu. Nếu dữ liệu có skewness lớn hơn 1 hoặc nhỏ hơn -1, có thể áp dụng các phép biến đổi như log, sqrt hoặc cbrt để giảm lệch.
- Áp dụng phép biến đổi (nếu cần) để đạt được phân phối gần phân phối chuẩn hơn.
- Sử dụng StandardScaler để chuẩn hóa dữ liệu bằng cách loại bỏ trung bình và chia độ lệch chuẩn của các biến.

Quá trình trên giúp cải thiện tính phân phối và đồng nhất dữ liệu số, từ đó tăng khả năng áp dụng các phương pháp thống kê và thuật toán học máy hiệu quả hơn trên dữ liệu này.

SK_ID_CURR	NAME_CONTRACT_TYPE	FLAG_OWN_CAR	FLAG_OWN_REALTY	NAME_EDUCATION_TYPE_Academic degree	NAME_EDUCATION_TYPE_Higher education	NAME_EDUCATION_TYPE_Incomplete higher	NAME_EDUCATION_TYPE_Lower secondary
100002	0	0	1	0	0	0	0
100003	0	0	0	0	1	0	0
100004	1	1	1	0	0	0	0
100006	0	0	1	0	0	0	0
100007	0	0	1	0	0	0	0

Hình 5.3. Một phần của bộ dữ liệu khi hoàn thành giai đoạn mã hóa

(Nguồn: Nhóm tác giả)

Dữ liệu cuối cùng sau khi mã hóa đặc biệt là mã hóa One-hot sẽ làm cho số lượng đặc trưng tăng lên đáng kể, cụ thể là 195 đặc trưng.

5.1.4. Xử lý dữ liệu mất cân bằng

Biến mục tiêu là 'TARGET' trong tập application train, với tỉ lệ 2 giá trị 0,1 là 91:9. Tỷ lệ mất cân bằng này cho thấy rằng lớp 0 (không xảy ra sự cố) chiếm đa số trong tập dữ liệu, trong khi lớp 1 (xảy ra sự cố) chỉ chiếm một phần nhỏ. Mất cân bằng dữ liệu như vậy có thể gây khó khăn trong việc xây dựng mô hình và đánh giá hiệu suất của nó. Quá trình xử lý mất cân bằng dữ liệu là một phần quan trọng trong quá trình xây dựng mô hình và cần được thực hiện cẩn thận để đảm bảo độ chính xác và độ tin cậy của mô hình.

Ở đây kỹ thuật oversampling được sử dụng là một kỹ thuật trong xử lý dữ liệu để giải quyết vấn đề mất cân bằng dữ liệu trong các bài toán phân loại. Kỹ thuật oversampling được sử dụng để tạo ra thêm các mẫu cho lớp thiểu số để tăng số lượng dữ liệu và cân bằng lại các lớp trong tập dữ liệu. Kỹ thuật oversampling có thể được thực hiện thông qua các phương pháp sao chép (duplicate), tạo mẫu (sampling) hoặc tăng cường dữ liệu (data augmentation), trong đó, chuyển đổi từ data cũ có thể được thực hiện thông qua việc sao chép các mẫu của lớp thiểu số nhiều lần để tạo ra các phiên bản mới của chúng. Sau đó, các phiên bản mới này sẽ được kết hợp với data cũ để tạo ra data mới có tổng cộng 565,372 dòng, cân bằng giữa các lớp với tỉ lệ lúc này là 1:1. Cuối cùng, với tập dữ liệu mới này, chúng ta có thể huấn luyện mô hình phân loại tốt hơn và đạt được độ chính xác cao hơn.

5.1.5. Tách bộ dữ liệu và lựa chọn đặc trưng

Đây là giai đoạn chuẩn bị cho quá trình đào tạo và kiểm tra mô hình. Việc đầu tiên là tách bộ dữ liệu thành 2 bộ dữ liệu mới là bộ dữ liệu đào tạo và tập dữ liệu kiểm tra với tỉ lệ 7:3.

Việc lựa chọn các trường dữ liệu được thực hiện bằng thuật toán Recursive Feature Elimination (RFE). Đây là một kỹ thuật lựa chọn đặc trưng trong học máy. Kỹ thuật được sử dụng để giảm số lượng đặc trưng trong dữ liệu, từ đó cải thiện độ chính xác của mô hình học máy, quá trình huấn luyện mô hình và tiết kiệm thời gian tính toán.

Như được thực hiện mã hóa ở bước trên làm cho số lượng cột gia tăng đáng kể vì vậy đây là một bước rất quan trọng và cần thiết. RFE khởi đầu bằng cách huấn luyện mô hình trên toàn bộ dữ liệu ban đầu và xác định mức độ quan trọng của từng đặc trưng. Sau đó, đặc trưng không quan trọng nhất được loại bỏ khỏi dữ liệu. Tiếp theo, quá trình này lặp đi lặp lại đến khi chỉ còn lại số lượng đặc trưng mong muốn hoặc tới khi độ chính xác mô hình không thay đổi nhiều.

Cụ thể, dữ liệu ban đầu có tông 195 đặc trưng sẽ loại bỏ đi các thuộc tính mã Id không cần thiết sẽ còn lại 188 đặc trưng. Sau đó, thuật toán RFE được sử dụng và đã chọn ra còn 97 đặc trưng phù hợp.

5.2. Mô hình dự đoán rủi ro thanh toán

Sau khi đã thực hiện các bước trên, dữ liệu được đưa vào mô hình học máy. Quá trình tiền xử lý dữ liệu là một bước cực kỳ quan trọng, góp phần quyết định độ chính xác của mô hình. Việc thực hiện chính xác quá trình tiền xử lý dữ liệu cho dataset Home Credit có thể giúp xây dựng được một mô hình học máy hiệu quả để đánh giá khả năng trả nợ của khách hàng.

5.2.1. Chỉ số đánh giá

Ma trận nhầm lẫn

		Predicted	
		Negative (**N**) -	Positive (**P**) +
Actual	Negative -	True Negatives (TN)	False Positives (FP) **Type I error**
	Positive +	False Negatives (FN) **Type II error**	True Positives (TP)

Hình 5.4. Ma trận nhầm lẫn

(Nguồn: https://www.nbshare.io/notebook/626706996/Learn-And-Code-Confusion-Matrix-With-Python/)

Ma trận nhầm lẫn (confusion matrix) là một công cụ quan trọng để đánh giá hiệu suất của mô hình dự đoán trong bài toán phân loại. Nó biểu thị số lượng dự đoán đúng và sai lệch cho từng lớp dữ liệu, giúp đánh giá chính xác khả năng phân loại của mô hình. Ma trận này gồm các thành phần như True Positive (TP), False Positive (FP), True Negative (TN) và False Negative (FN). Bằng cách xem xét các giá trị TP, FP, TN và FN, ta có thể tính toán các độ đo như accuracy, precision, recall và F1 score để đánh giá và so sánh hiệu suất của các mô hình khác nhau. Ma trận nhầm lẫn cung cấp thông tin chi tiết về hiệu suất của mô hình phân loại và giúp chúng ta đánh giá chính xác khả năng phân loại của mô hình dự đoán.

Accuracy, Precision, Recall and F1 Score

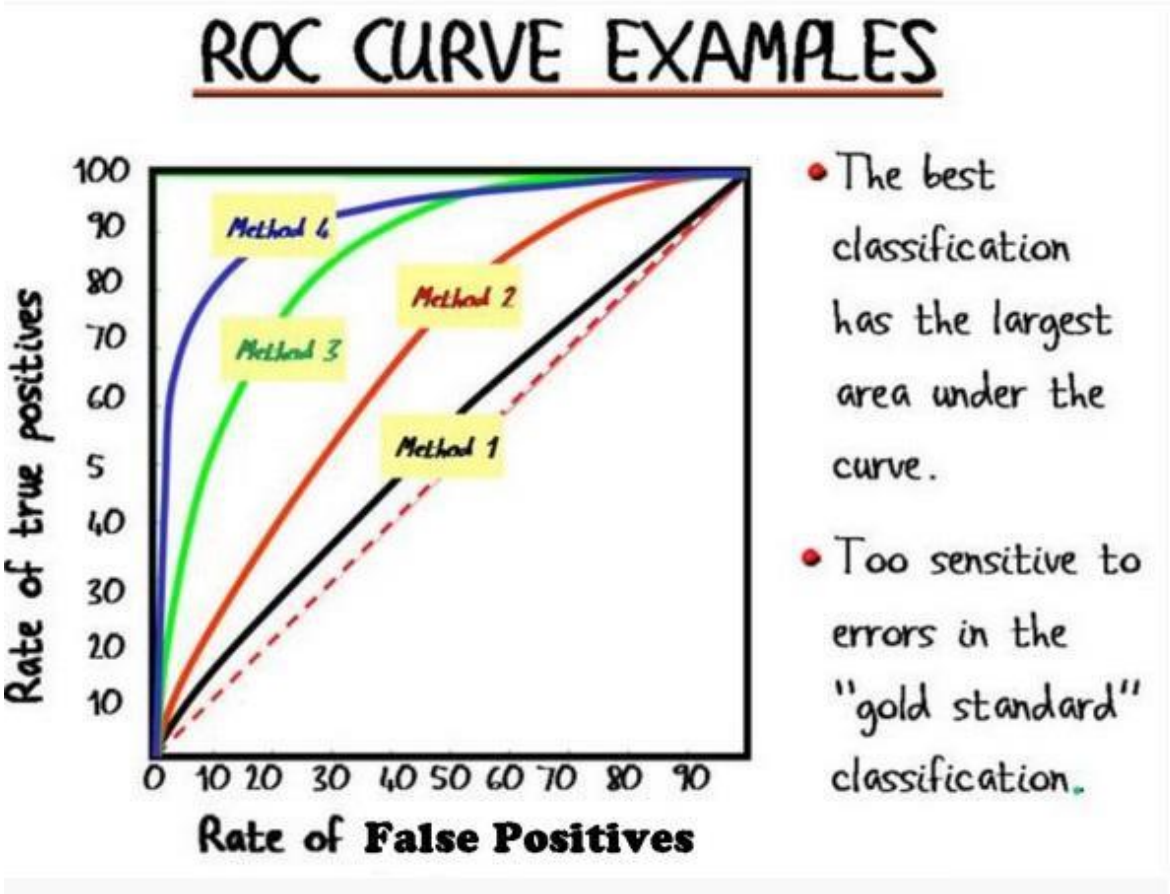
Accuracy, precision, recall và F1 score là các độ đo quan trọng để đánh giá hiệu suất của mô hình dự đoán trong bài toán phân loại. Đây là những độ đo được tính toán từ ma trận nhầm lẫn (confusion matrix) để cung cấp thông tin chi tiết về khả năng phân loại chính xác và độ phủ của mô hình trên từng lớp dữ liệu.

- Accuracy (độ chính xác) đo lường tỷ lệ phần trăm các mẫu được dự đoán đúng trên tổng số mẫu. Nó cho biết mức độ chính xác của mô hình toàn bộ và thể hiện khả năng dự đoán đúng cả các mẫu positive và negative.
- Precision (độ chính xác của dương tính) là tỷ lệ giữa số lượng dự đoán đúng là positive (TP) so với tổng số dự đoán positive (TP + FP). Độ đo này đánh giá khả năng của mô hình trong việc đưa ra các dự đoán positive chính xác.
- Recall (độ phủ) là tỷ lệ giữa số lượng dự đoán đúng là positive (TP) so với tổng số mẫu thực tế positive (TP + FN). Recall đo lường khả năng của mô hình trong việc tìm ra tất cả các mẫu positive có thể.
- F1 score là một phép kết hợp của precision và recall, được tính bằng công thức 2 * (Precision * Recall) / (Precision + Recall). Độ đo này kết hợp cả khả năng dự đoán chính xác positive và khả năng tìm ra tất cả các mẫu positive, giúp đánh giá tổng thể hiệu suất của mô hình.

Các độ đo accuracy, precision, recall và F1 score cung cấp thông tin quan trọng về hiệu suất của mô hình phân loại. Chúng giúp chúng ta đánh giá và so sánh khả năng phân loại của các mô hình khác nhau và lựa chọn mô hình phù hợp với yêu cầu của bài toán.

AUC

TP.HCM, ngày 15 tháng 6 năm 2023



Hình 5.5. Ví dụ về ROC Curve

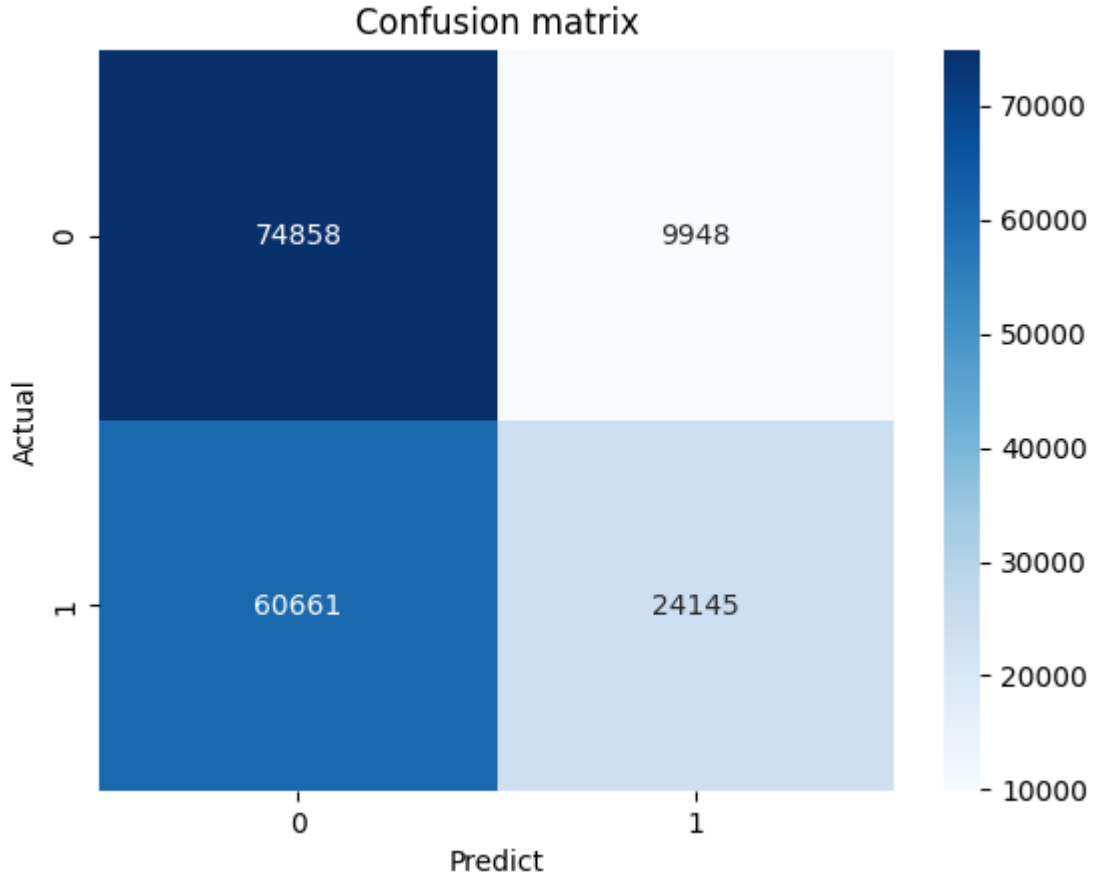
(Nguồn: https://hrngok.github.io/posts/roc%20curve%20%26%20auc/)

AUC (Area Under the ROC Curve) là một độ đo quan trọng được sử dụng để đánh giá hiệu suất của mô hình dự đoán trong bài toán phân loại. Nó được tính dựa trên đường cong ROC (Receiver Operating Characteristic), mô tả sự biến thiên của tỷ lệ True Positive Rate (TPR) và False Positive Rate (FPR) khi ngưỡng phân loại thay đổi. Đường cong ROC là một đồ thị biểu diễn mối quan hệ giữa TPR và FPR, trong đó TPR là tỷ lệ dự đoán đúng positive và FPR là tỷ lệ dự đoán sai negative. AUC là diện tích nằm dưới đường cong ROC, và giá trị AUC cho thấy khả năng phân loại của mô hình dự đoán. AUC là một độ đo phổ biến và quan trọng trong việc đánh giá hiệu suất của mô hình phân loại, đặc biệt khi mất cân bằng giữa các lớp dữ liệu. Giá trị AUC nằm trong khoảng từ 0 đến 1, với AUC = 0.5 cho thấy mô hình dự đoán ngẫu nhiên, và AUC = 1 cho thấy mô hình có khả năng phân loại hoàn hảo. AUC cung cấp thông tin về khả năng phân loại của mô hình và độ chắc chắn của dự đoán. Một giá trị

AUC cao cho thấy mô hình có khả năng phân loại tốt hơn, với TPR cao và FPR thấp. Điều này cho thấy mô hình có khả năng tìm ra các mẫu positive và đồng thời giảm sai lệch positive. AUC cũng cho phép so sánh và lựa chọn mô hình phân loại phù hợp với yêu cầu cụ thể của bài toán. Khi so sánh hai mô hình, mô hình có AUC cao hơn được coi là có hiệu suất phân loại tốt hơn. Với tỷ lệ False Positive và False Negative khác nhau, AUC cung cấp một cái nhìn tổng quan về hiệu suất của mô hình phân loại và giúp chúng ta hiểu rõ hơn về khả năng phân loại của mô hình trong bài toán phân loại.

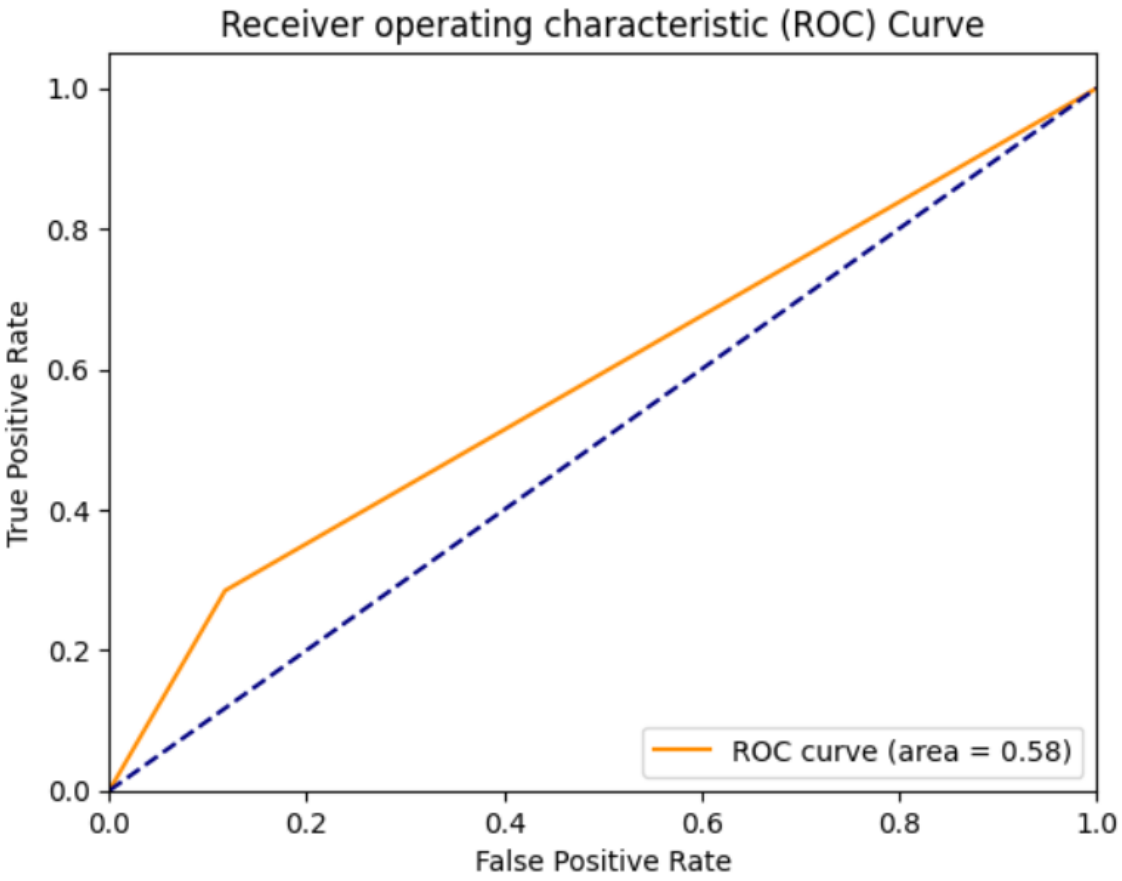
5.2.2. Logistics regression

Mô hình Logistics regression cho kết quả dự đoán được thể hiện qua Bảng. Thông qua ma trận nhầm lẫn, ta thấy được mô hình dự đoán khá tốt với nhóm khách hàng an toàn nhưng còn hạn chế trong khả năng dự đoán nhóm khách hàng có rủi ro thanh toán mà đây lại là đối tượng mục tiêu của bài toán.



Hình 5.6. Ma trận nhầm lẫn dự đoán rủi ro thanh toán bằng Logistics Regression

(Nguồn: Nhóm tác giả)



Hình 5.7. AUC kết quả dự đoán rủi ro thanh toán bằng Logistics Regression

(Nguồn: Nhóm tác giả)

Như được thể hiện trên hình AUC có kết quả là 0.58. Điều này có ý nghĩa là mô hình có khả năng phân loại các mẫu Positive và Negative không tốt.

	precision	recall	f1-score	support
0	0.55	0.88	0.68	84806
1	0.71	0.28	0.41	84806
accuracy			0.58	169612

macro avg	0.63	0.58	0.54	169612
weighted avg	0.63	0.58	0.54	169612

Bảng 5.1. Kết quả dự đoán rủi ro thanh toán bằng Logistics Regression

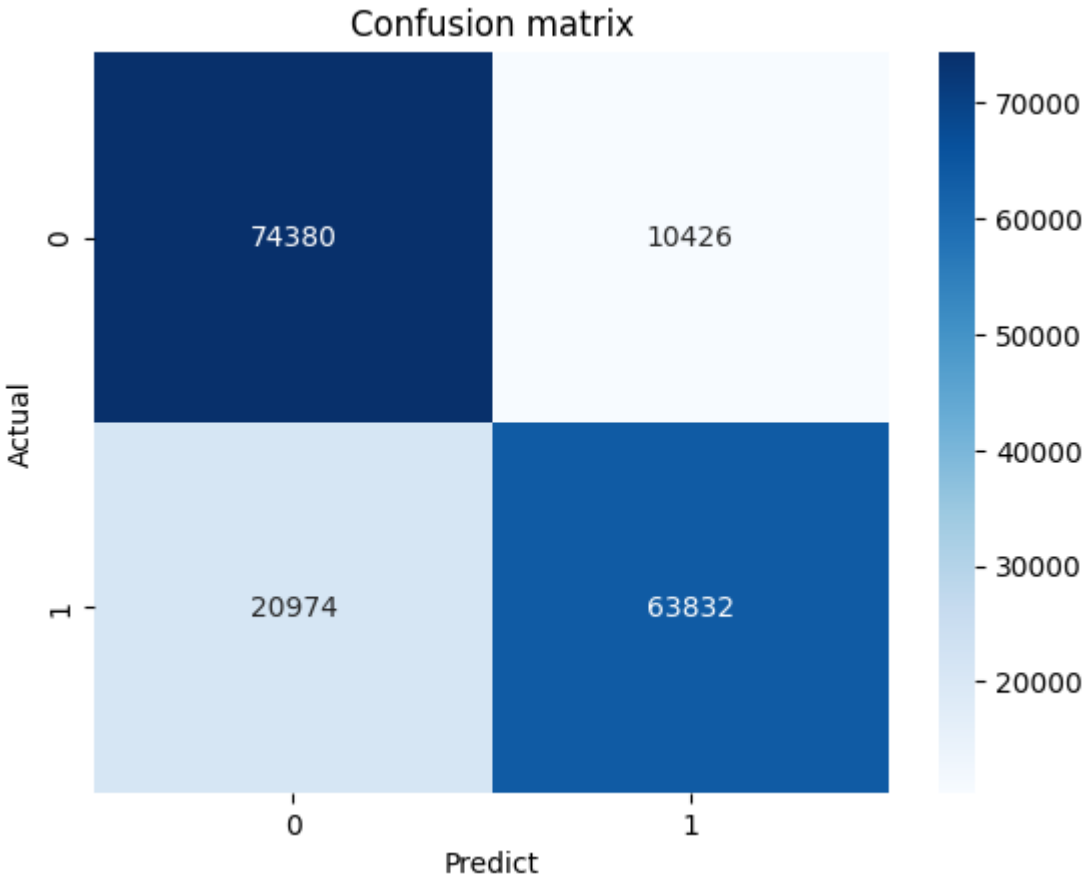
- Precision cho lớp 0 là 0.55 và cho lớp 1 là 0.71. Precision cho lớp 1 có giá trị thấp, cho thấy mô hình có xu hướng phân loại nhầm nhiều mẫu Negative thành Positive.
- Recall cho lớp 0 là 0.88 và cho lớp 1 là 0.28. Kết quả cho thấy mô hình có khả năng bỏ sót nhiều mẫu Positive.
- F1-score cho lớp 0 là 0.68 và cho lớp 1 là 0.41. Giá trị F1-score thấp cho thấy mô hình không đạt được sự cân bằng tốt giữa precision và recall.

Tổng quan, dựa vào báo cáo đánh giá này, mô hình có kết quả tương đối cho lớp 0 (không rủi ro) với độ chính xác và độ truy tìm cao. Tuy nhiên, mô hình có kết quả kém hơn cho lớp 1 (rủi ro) với độ chính xác và độ truy tìm thấp.

Kết quả đánh giá của mô hình cho thấy hiệu suất phân loại cần được cải thiện. Mô hình có xu hướng phân loại nhầm nhiều mẫu Negative thành Positive (precision thấp) và bỏ sót nhiều mẫu Positive thực tế (recall thấp).

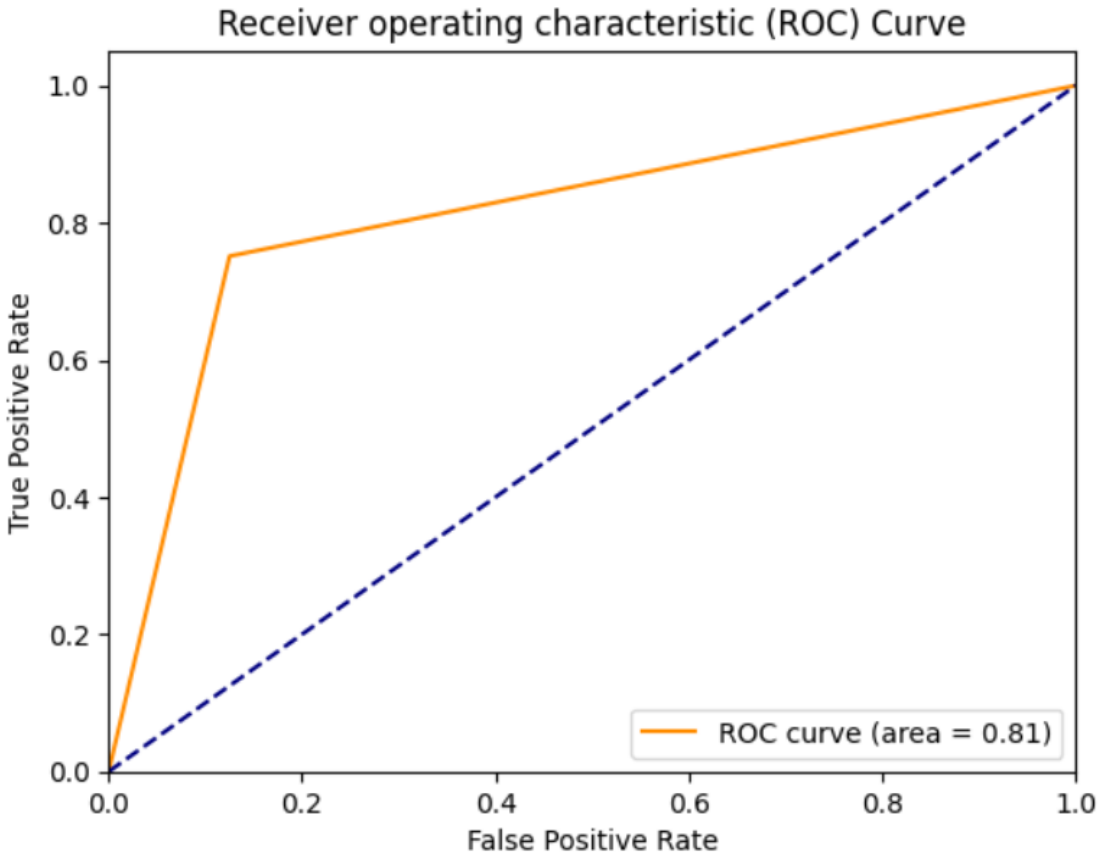
5.2.3. Random Forest

Mô hình Random Forest cho kết quả cải thiện hơn khi đối tượng rủi ro được phát hiện tốt hơn được thể hiện ở ma trận nhầm lẫn ở bên dưới.



Hình 5.8. Ma trận nhầm lẫn dự đoán rủi ro thanh toán bằng Random Forest

(Nguồn: Nhóm tác giả)



Hình 5.9. AUC kết quả dự đoán rủi ro thanh toán bằng Random Forest

(Nguồn: Nhóm tác giả)

Với giá trị AUC = 0.81, mô hình Random Forest cho thấy khả năng phân loại rủi ro tín dụng của khách hàng Home Credit là tương đối tốt. Mô hình có khả năng tách biệt tốt giữa các mẫu thuộc hai lớp khác nhau và có khả năng đưa ra dự đoán chính xác.

Để đánh giá chi tiết hơn thì bảng báo cáo kết quả được thể hiện ở bên dưới.

	precision	recall	f1-score	support
0	0.78	0.88	0.83	84806
1	0.86	0.75	0.80	84806

accuracy			0.81	169612
macro avg	0.82	0.81	0.81	169612
weighted avg	0.82	0.81	0.81	169612

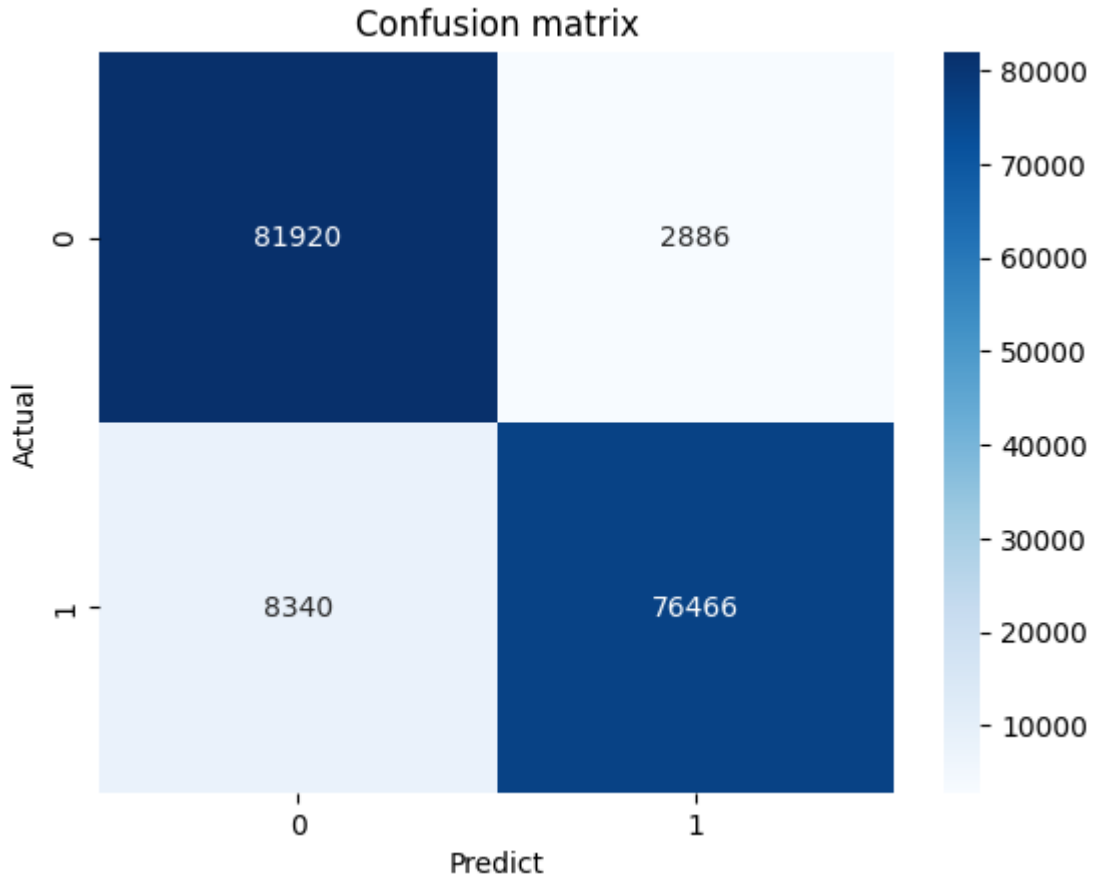
Bảng 5.2. Kết quả dự đoán rủi ro thanh toán bằng Random Forest

- Precision (Độ chính xác): Trong trường hợp lớp 0 độ chính xác là 0.78, trong khi đó trong trường hợp lớp 1 độ chính xác là 0.86. Điều này cho thấy mô hình có khả năng phân loại chính xác các mẫu thuộc các lớp khác nhau.
- Recall (Độ bao phủ): Trong trường hợp lớp 0 độ bao phủ là 0.88, trong khi đó trong trường hợp lớp 1 độ bao phủ là 0.75. Kết quả cho thấy mô hình có độ bao phủ tương đối cao.
- F1-score: Trong trường hợp lớp 0, F1-score là 0.83, trong khi đó trong trường hợp lớp 1, F1-score là 0.80. Hiệu suất của mô hình khá cân bằng giữa độ chính xác và độ bao phủ.

Tuy nhiên, mục tiêu bài toán dự đoán rủi ro mà chỉ số recall ở class 1 chưa cao, nghĩa là có mô hình còn hạn chế ở đanh giá hồ sơ có rủi ro thanh toán trở thành hồ sơ an toàn. Điều này có thể gây ra tổn thất cho doanh nghiệp.

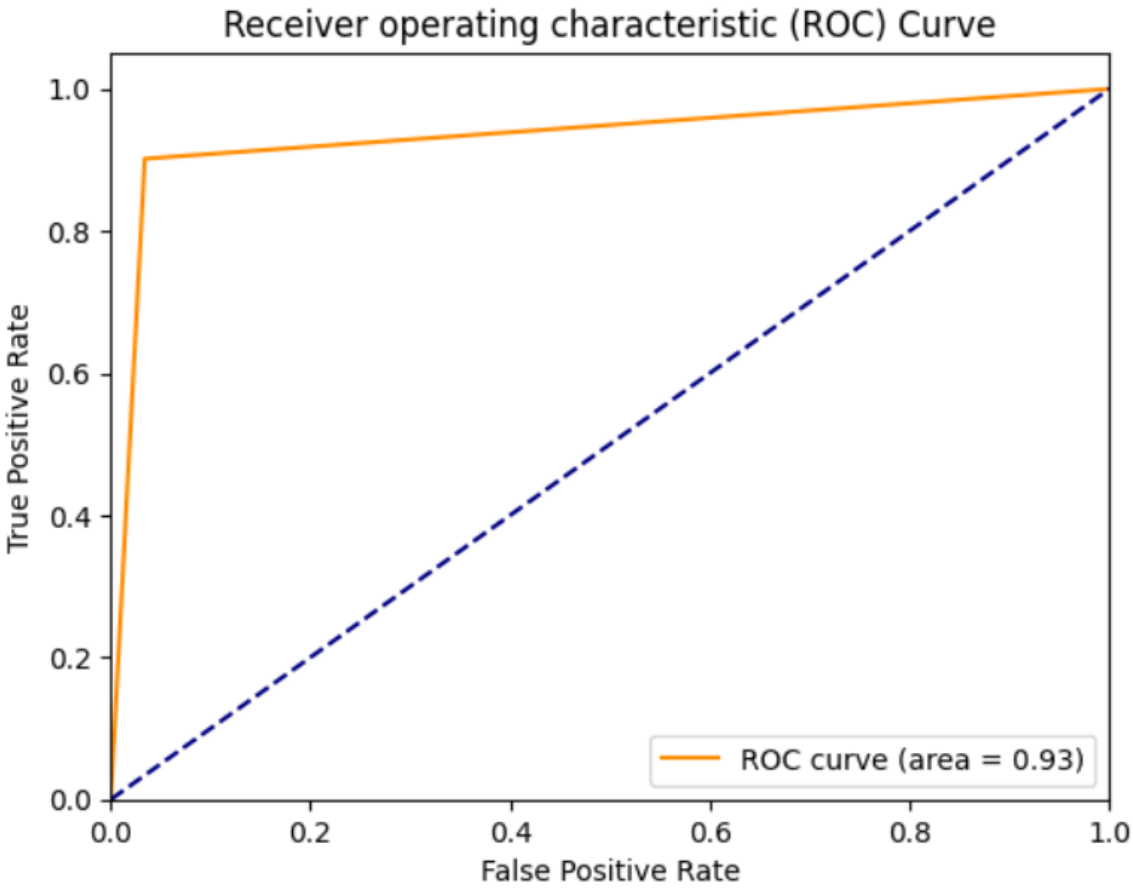
5.2.4. XGBoost

XGBoost cho kết quả dự đoán cải thiện đáng kể so với mô hình còn lại. Điều này được quan sát qua kết quả ma trận nhầm lẫn ở **Bảng** .



Hình 5.10. Ma trận nhầm lẫn dự đoán rủi ro thanh toán bằng XGBoost

(Nguồn: Nhóm tác giả)



Hình 5.11. AUC kết quả dự đoán rủi ro thanh toán bằng XGBoost

(Nguồn: Nhóm tác giả)

Với giá trị AUC = 0.93, mô hình XGBoost cho thấy khả năng phân loại rủi ro tín dụng của khách hàng Home Credit là rất tốt. Mô hình có khả năng tách biệt cao giữa các mẫu thuộc hai lớp khác nhau và có khả năng đưa ra dự đoán chính xác.
Để đánh giá cụ thể xem xét báo cáo kết quả ở Bảng phía dưới.

	precision	recall	f1-score	support
0	0.91	0.97	0.94	84806
1	0.96	0.90	0.93	84806
accuracy			0.93	169612

macro avg	0.94	0.93	0.93	169612
weighted avg	0.94	0.93	0.93	169612

Bảng 5.3. Kết quả dự đoán rủi ro thanh toán bằng XGBoost

Độ đo precision, recall và f1-score cho lớp 0 (không rủi ro) và lớp 1 (rủi ro) đều rất cao. Điều này cho thấy mô hình có khả năng đưa ra dự đoán chính xác và có độ chính xác cao trong việc phân loại khách hàng vào các nhóm rủi ro và không rủi ro. Dựa vào các đánh giá ở trên cho thấy rằng mô hình XGBoost được sử dụng để dự đoán rủi ro thanh toán là phù hợp nhất, cho kết quả chính xác nhất.

Vì vậy, XGBoost được dùng để dự đoán tệp khách hàng mới để sàng lọc những hồ sơ khách hàng có uy tín và ít rủi ro thanh toán.

CHƯƠNG 6: KẾT QUẢ VÀ THẢO LUẬN

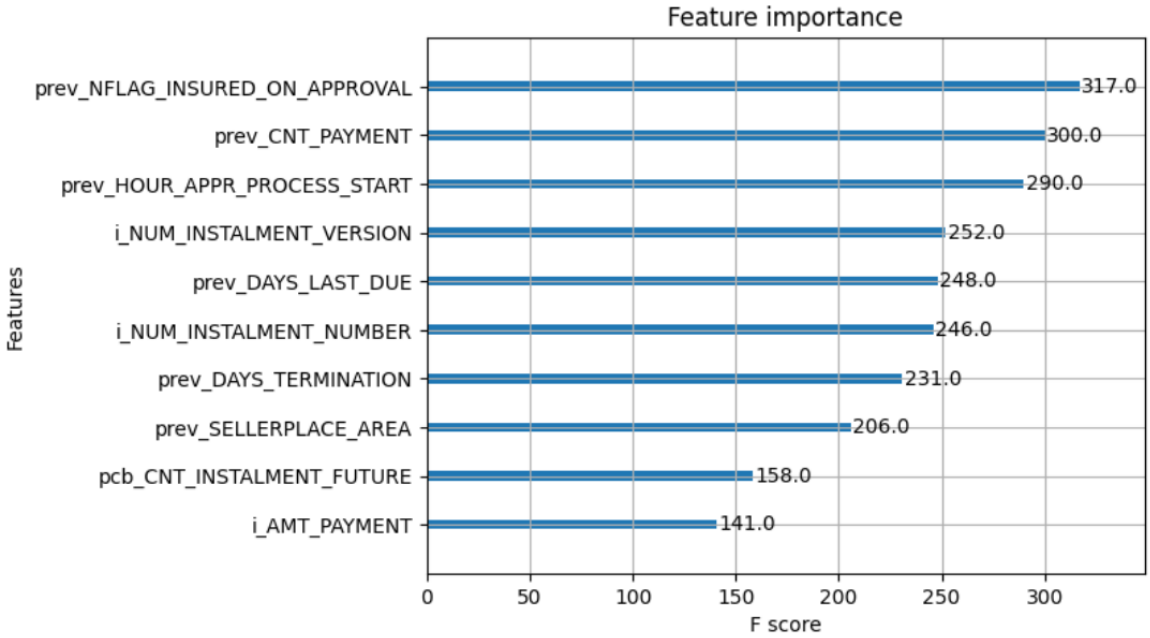
Quá trình dự đoán rủi ro thanh toán là một trong những vấn đề quan trọng trong các hoạt động tài chính. Tính độ chính xác của mô hình dự đoán rủi ro thanh toán là yếu tố rất quan trọng để đảm bảo tính bảo mật và thực hiện các quyết định tài chính chính xác hơn. Ở đây, mô hình XGBoost, Logistics Regression và Random Forest đã được đánh giá và so sánh với nhau. Kết quả cho thấy rằng XGBoost là mô hình cho kết quả dự đoán chính xác và vượt trội hơn hẳn so với mô hình còn lại.

Đặc biệt, XGBoost đạt được hiệu suất cao với chỉ số Accuracy và F1-Score. Với những chỉ số này đạt mức cao, mô hình XGBoost là một giải pháp lý tưởng cho việc ứng dụng dự đoán và đánh giá các hồ sơ đăng ký vay. Vì vậy, mô hình XGBoost có thể cải thiện tính chính xác trong dự đoán rủi ro thanh toán, đồng thời tăng tính bảo mật và đáng tin cậy trong các quyết định tài chính. Các kết quả thực nghiệm này đã chứng minh rằng XGBoost là một lựa chọn tốt để áp dụng trong các hoạt động tài chính phức tạp.

	precision	recall	f1-score	accuracy	AUC
Logictics Regression	0.63	0.58	0.54	0.58	0.58
Random Forest	0.82	0.81	0.81	0.81	0.81
XGBoost	0.94	0.93	0.93	0.93	0.93

Bảng 6.1. So sánh kết quả dự đoán giữa các mô hình

Ngoài ra, qua mô hình dự đoán rủi ro thanh toán, các biến tác động nhiều nhất đến khả năng thanh toán cũng được phát hiện. Mức độ ảnh hưởng của 10 yếu tố có tác động nhất đến mô hình được thể hiện ở **Hình** bên dưới.



Hình 6.1. Mức độ tác động của các biến

(Nguồn: Nhóm tác giả)

Dữ liệu "NFLAG_INSURED_ON_APPROVAL" là một dữ liệu được đánh giá là quan trọng nhất trong mô hình đánh giá rủi ro vay tín dụng của công ty HomeCredit. Nó được xem là dữ liệu quan trọng nhất vì nó liên quan đến yếu tố an ninh trong giao dịch vay của khách hàng. Dữ liệu này thể hiện mối quan hệ giữa việc khách hàng có được bảo hiểm hay không và xác nhận từ ngân hàng về khoản vay. Giải thích cho tầm quan trọng cao hơn của dữ liệu này so với các dữ liệu khác có thể là do khi khách hàng đã được bảo hiểm và xác nhận về khoản vay, tỷ lệ rủi ro của việc không trả tiền có thể giảm xuống trong tương lai. Điều này cho thấy rằng việc có bảo hiểm và xác nhận từ ngân hàng có thể làm giảm nguy cơ mất nợ và tăng tính bảo mật trong quá trình giao dịch vay tín dụng. Do đó, việc quan tâm và đánh giá kỹ lưỡng yếu tố này có thể giúp công ty HomeCredit đưa ra các quyết định vay tín dụng một cách an toàn và có hạn chế rủi ro.

Ngoài ra, các dữ liệu có tác động lớn đến mô hình đánh giá rủi ro vay tín dụng trong công ty chủ yếu là các dữ liệu từ các đơn đăng ký vay trước đó của khách hàng

(thông tin từ bảng previous application). Các biến này cung cấp cho mô hình một cái nhìn tổng quan về khả năng và lịch sử thanh toán nợ của khách hàng, và có vai trò cực kỳ quan trọng trong dự đoán rủi ro thanh toán hiện tại và trong tương lai của khách hàng. Điều này cực kỳ quan trọng đối với mô hình dự đoán rủi ro thanh toán vì chúng giúp dự đoán khả năng thanh toán nợ hiện tại và trong tương lai của khách hàng. Cụ thể, dữ liệu "prev_CNT_PAYMENT" liên quan đến số lượng khoản thanh toán cho khoản vay trước đó. Mức độ ảnh hưởng lớn của dữ liệu này cho thấy số lượng khoản thanh toán có thể có tác động quan trọng đến quá trình đánh giá rủi ro vay. Số lượng khoản thanh toán có thể phản ánh khả năng và tình trạng thanh toán của khoản vay. Dữ liệu "prev_SELLERPLACE_AREA" liên quan đến diện tích của nơi mua hàng trong giao dịch trước đó. Mức độ ảnh hưởng lớn của biến này cho thấy diện tích của nơi mua hàng có thể có tác động đáng kể đến quá trình đánh giá rủi ro vay. Diện tích này có thể liên quan đến loại hình và quy mô của giao dịch mua hàng. Tóm lại, thông qua việc xem xét các biến quan trọng như "prev_CNT_PAYMENT" và "prev_SELLERPLACE_AREA", ta có thể nhận thấy tầm quan trọng của các yếu tố liên quan đến lịch sử vay trước đó, bao gồm số lượng khoản thanh toán, diện tích nơi mua hàng và số tiền trả hàng tháng. Các yếu tố này cung cấp thông tin quan trọng về khả năng thanh toán nợ và tình trạng tài chính của khách hàng, giúp công ty HomeCredit đánh giá và quản lý rủi ro tín dụng một cách hiệu quả.

Ngoài ra, các biến liên quan đến thời hạn thanh toán cũng xuất hiện nhiều trong top 10 biến có tác động đến mô hình. Vì vậy, việc xem xét kỹ hồ sơ và lịch sử thanh toán của khách hàng là rất quan trọng. Điều này giúp đưa ra một đánh giá chính xác hơn về khả năng thanh toán nợ của khách hàng trong quá khứ và hiện tại. Nếu khách hàng đã từng trễ hạn hoặc không thanh toán nợ đúng hạn trong các khoản vay trước đó, khả năng cao họ sẽ có xu hướng trễ hạn trong tương lai. Cụ thể, dữ liệu "prev_HOUR_APPR_PROCESS_START" liên quan đến thời gian trong ngày xác nhận đơn vay trước đó. Mức độ ảnh hưởng lớn của dữ liệu này cho thấy thời gian trong ngày có thể có tác động đáng kể đến quá trình đánh giá rủi ro vay. Thời gian xác nhận đơn vay có thể liên quan đến khả năng thanh toán và tính chính xác trong quá trình xác nhận thông tin vay. Tiếp theo, dữ liệu "prev_DAYS_LAST_DUE" liên quan đến số

ngày còn lại cho đến kỳ hạn trả góp cuối cùng của khoản vay trước đó. Mức độ ảnh hưởng lớn của dữ liệu này cho thấy số ngày còn lại có thể có tác động đáng kể đến quá trình đánh giá rủi ro vay. Số ngày còn lại có thể phản ánh khả năng và tính chính xác trong việc trả góp của khoản vay. Ngoài ra, dữ liệu "prev_DAYS_TERMINATION" liên quan đến số ngày từ kỳ hạn trả góp cuối cùng cho đến khi kết thúc khoản vay trước đó. Mức độ ảnh hưởng lớn của dữ liệu này cho thấy số ngày từ kỳ hạn trả góp cuối cùng đến khi kết thúc có thể có tác động đáng kể đến quá trình đánh giá rủi ro vay. Số ngày này có thể liên quan đến tính chính xác và thời gian của quá trình trả góp. Cuối cùng "prev_HOUR_APPR_PROCESS_START" là dữ liệu liên quan đến thời gian trong ngày xác nhận đơn vay trước đó. Mức độ ảnh hưởng lớn của dữ liệu này cho thấy thời gian trong ngày có thể có tác động đáng kể đến quá trình đánh giá rủi ro vay. Thời gian xác nhận đơn vay có thể liên quan đến khả năng thanh toán và tính chính xác trong quá trình xác nhận thông tin vay. Tóm lại, việc xem xét kỹ các biến liên quan đến thời hạn thanh toán giúp đánh giá chính xác hơn về khả năng thanh toán nợ của khách hàng trong quá khứ và hiện tại. Điều này là rất quan trọng để công ty có thể quản lý rủi ro tín dụng một cách hiệu quả.

Thông tin chi tiết về các khoản trả góp của khoản vay trước đó trong bảng installment cũng có tác động quan trọng đến mô hình. Đầu tiên, dữ liệu "i_NUM_INSTALMENT_VERSION" liên quan đến phiên bản số kỳ hạn trả góp cho khoản vay trước đó. Mức độ ảnh hưởng lớn của biến này cho thấy phiên bản số kỳ hạn trả góp có vai trò quan trọng trong quá trình đánh giá rủi ro vay. Phiên bản số kỳ hạn trả góp có thể ảnh hưởng đến lịch trình và điều kiện trả góp của khoản vay. Tiếp theo, dữ liệu "i_NUM_INSTALMENT_NUMBER" liên quan đến số lượng kỳ hạn trả góp cho khoản vay trước đó. Mức độ ảnh hưởng lớn của biến này cho thấy số lượng kỳ hạn trả góp có vai trò quan trọng trong quá trình đánh giá rủi ro vay. Số lượng kỳ hạn trả góp có thể ảnh hưởng đến thời gian và tần suất trả góp của khoản vay. Ngoài ra, dữ liệu "i_AMT_PAYMENT" liên quan đến số tiền đã thanh toán trong kỳ trước đó của khoản vay. Mức độ ảnh hưởng lớn của biến này cho thấy số tiền thanh toán có thể có tác động đáng kể đến quá trình đánh giá rủi ro vay. Số tiền thanh toán có thể liên quan đến khả năng thanh toán và độ tin cậy tài chính của khách hàng. Tóm lại, thông tin chi

tiết về các khoản trả góp của khoản vay trước đó trong bảng installment đóng vai trò quan trọng trong mô hình đánh giá rủi ro vay. Phiên bản số kỳ hạn trả góp và số lượng kỳ hạn trả góp đều có thể ảnh hưởng đáng kể đến quá trình trả góp và rủi ro tín dụng.

Thông tin về các khoản vay POS (điểm bán hàng) và tiền mặt trước đó mà người đăng ký đã có với Home Credit cũng đóng vai trò quan trọng. Dữ liệu "pcb_CNT_INSTALMENT_FUTURE" liên quan đến số lượng kỳ hạn trả góp trong tương lai cho khoản vay trước đó. Mức độ ảnh hưởng lớn của biến này cho thấy số lượng kỳ hạn trả góp trong tương lai có tác động quan trọng đến quá trình đánh giá rủi ro vay. Số lượng này có thể ảnh hưởng đến lịch trình và điều kiện trả góp của khoản vay trong tương lai. Nếu khách hàng có số kỳ hạn trả góp còn lại ít, tức là gần hoàn thành việc trả nợ, rủi ro tín dụng có thể giảm. Tóm lại, thông tin về các khoản vay POS và tiền mặt trước đó với Home Credit đóng vai trò quan trọng trong mô hình đánh giá rủi ro vay. Số lượng kỳ hạn trả góp trong tương lai của khoản vay trước đó có thể ảnh hưởng đáng kể đến quá trình trả góp và rủi ro tín dụng.

Mô hình XGBoost được dùng với tập dữ liệu applications test để phân loại hồ sơ, kết quả được hiển thị ở bảng dưới.

	0 (khách hàng ít rủi ro)	1 (khách hàng rủi ro)
XGBoost	48087 hồ sơ	657 hồ sơ

Bảng 6.2. Kết quả dự đoán rủi ro thanh toán bằng XGBoost với tập application_test

CHƯƠNG 7: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Đầu tiên, là việc tìm hiểu về nghiệp vụ tài chính cũng như cơ sở dữ liệu của công ty như một vài trò của Data Analyst nhóm đã có cái nhìn rõ hơn về các công việc và nhiệm vụ ở vị trí này qua sự hướng dẫn của anh Thạch, Recovery Data Analytics Team Leader. Nhóm đã thực hiện các thống kê mô tả và thống kê suy luận để hiểu về thông tin dữ liệu như data storytelling, từ đó xem xét các yếu tố tác động đến mô hình để đưa ra các lời khuyên bổ ích giúp cho nhà quản trị có thể đưa ra các quyết định cho nhân viên duyệt hồ sơ vay mang tính rủi ro.

Về mục tiêu tìm hiểu kiến thức tài chính và đánh giá hồ sơ vay nợ, nhóm đã có được kiến thức cơ bản. nhóm đã tìm hiểu được những kiến thức cơ bản về tài chính và đánh giá hồ sơ vay nợ. Đó là các thông tin liên quan đến khoản vay, các thuật ngữ trong ngành tài chính và các sản phẩm của một công ty tài chính. Nhóm đã được giải thích các khái niệm và tiêu chuẩn đánh giá về khả năng thanh toán của khách hàng, các chỉ tiêu để đánh giá rủi ro của khoản vay và cách tính toán lãi suất vay. Đồng thời, nhóm đã được giới thiệu với các sản phẩm tài chính của công ty tài chính, từ đó hiểu được cách áp dụng các sản phẩm tài chính và những tiêu chuẩn để đánh giá khách hàng có thể đủ điều kiện vay tiền. Ngoài ra, nhóm còn tìm hiểu và đưa ra các giải pháp cho việc quản lý và giảm thiểu rủi ro trong hoạt động tài chính tổng thể của công ty.

Sau khi tìm hiểu các kiến thức và kỹ năng cơ bản trong ngành tài chính và có được cái nhìn tổng quan về các sản phẩm tài chính của công ty, và quan trọng hơn hết là hiểu tổng quan về bộ dataset được thực hiện nhóm đã tiếp tục áp dụng và phát triển các kiến thức về phân tích dữ liệu và máy học để đưa ra mô hình dự đoán rủi ro và phân nhóm hồ sơ khách hàng. Nhóm đã xây dựng một số mô hình máy học phổ biến trong đó là Logistics Regression, Random Forest và XGBoost. Trong đó sử dụng thuật toán XGBoost được đánh giá là mô hình có khả năng phát hiện rủi ro với độ tin cậy cao qua ma trận nhầm lẫn, chỉ số AUC với 93%, báo cáo đánh giá đặc biệt là F1-Score với 93% cho thấy mô hình khá cân bằng giữa độ chính xác và độ bao phủ. Từ đó mô hình XGBoost được sử dụng để dự đoán về rủi ro của hồ sơ vay trong tập dữ liệu Application Test. Mô hình được đánh giá là có tính ứng dụng, giúp công ty tài chính

TNHH MTV Home Credit Việt Nam quản lý rủi ro tín dụng và tối đa hóa lợi nhuận khi cung cấp dịch vụ tài chính cho khách hàng.

Thông qua việc huấn luyện và tinh chỉnh các mô hình này, nhóm cũng đã phát hiện ra các yếu tố tác động đến rủi ro thanh toán của khách hàng dựa trên đặc trưng ảnh hưởng lớn đối với mô hình dự đoán XGBoost. Điều này giúp cho nhóm có những kiến thức cần thiết để đưa ra các giải pháp và khuyến nghị cho công ty. Top 10 đặc trưng ảnh hưởng đến mô hình nhiều nhất được lựa chọn và đánh giá. Đây là bước vô cùng quan trọng để cung cấp insight giúp nhà hỗ trợ ra quyết định, chuyên gia về lĩnh vực tài chính có thể dùng làm tài liệu để đánh giá về nghiệp vụ tài chính rủi ro có cơ sở. Từ đó giúp cho doanh nghiệp có thể phát hiện ra các hồ sơ rủi ro một cách dễ dàng hơn, tăng cường hiệu quả của hoạt động kinh doanh và cải thiện trải nghiệm khách hàng của công ty tài chính nói chung và Home Credit nói riêng.

Tuy nhiên, mặc dù nhóm đã cố gắng hoàn thành chỉnh chu nhất có thể nhưng nhóm vẫn còn một số hạn chế. Đầu tiên là hạn chế về mặt thời gian và kiến thức nền, trong khoảng thời gian ngắn nhóm vẫn chưa thể hiểu sâu được phân tích nghiệp vụ tài chính khi nhóm phải tìm hiểu từ đầu với những định nghĩa cơ bản nhất về tài chính. Và cũng vì thế mà sản phẩm cuối nhóm chỉ có thể đưa ra những insight cơ bản, chưa có phân tích sâu hơn về đặc trưng như một chuyên gia để có thể đề xuất cụ thể cho sự phát triển của doanh nghiệp. Bên cạnh đó, còn có rất nhiều phương pháp để xây dựng mô hình dự đoán rủi ro vay tài chính. Trong tương lai nhóm sẽ đầu tư kiến thức tài chính, cũng như xây dựng nhiều phương pháp khác nhau để có thể đánh giá theo nhiều chiều khác nhau, có sự so sánh khách quan để thấy được nhiều tác động khác nhau giúp cho việc xây dựng mô hình chính xác, khách quan hơn.

TP.HCM, ngày 15 tháng 6 năm 2023

TÀI LIỆU THAM KHẢO

Home Credit Vietnam. (2021). About us. Retrieved June 10, 2023, from https://www.homecredit.vn/about-us.

Chen, H. (2019). Credit risk assessment and management for SMEs based on big data: A case study in China. Information Processing & Management, 56(4), 1600-1614. doi: 10.1016/j.ipm.2019.03.003

Chu, T. W., Zhang, P., Wang, X., & Yang, X. (2020). Financial risk management of supply chain finance based on the credit risk of core enterprises. Journal of Applied Accounting and Finance, 2(1), 1-15. doi: 10.22158/jaaf.v2n1p1

Dinh, T. V., Nguyen, T. H., Nguyen, T. T. H., & Nguyen, N. T. (2020). A hybrid method for credit risk evaluation based on machine learning and fuzzy logic. Expert Systems with Applications, 139, 112831. doi: 10.1016/j.eswa.2019.112831

Wongnaa, Camillus & Awunyo-Vitor, Dadson. (2013). Factors Affecting Loan Repayment Performance Among Yam Farmers in the Sene District, Ghana. Agris On-line Papers in Economics and Informatics. 5. 111-122.

Muriithi, Jane & Waweru, Kennedy & Muturi, Willy. (2016). Effect of Credit Risk on Financial Performance of Commercial Banks Kenya. IOSR Journal of Economics and Finance. 07. 72-83. 10.9790/5933-0704017283.

Schneider, G. P. (2011). Electronic Commerce 9th edition, 9 penyunt. *United States of America: Cengage Learning*, 4.

Sari, J. N., Nugroho, L. E., Ferdiana, R., & Santosa, P. I. (2016). Review on customer segmentation technique on ecommerce. *Advanced Science Letters*, *22*(10), 3018-3022.

Machinelearningcoban(2017). Bài 10: Logistic Regression. Retrieved December 16, 2022. from: https://machinelearningcoban.com/2017/01/27/ logisticregression/.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

What is Random Forest? | IBM. (n.d.-b). https://www.ibm.com/topics/random-forest

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

Introduction to random forest in machine learning (no date) *Section*, tại https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/ (Ngày truy cập cuối 14/06/2023).

Decision tree (2023) GeeksforGeeks, tại https://www.geeksforgeeks.org/decision-tree/ (Ngày truy cập cuối 14/06/2023).

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of clinical epidemiology, 110, 12-22.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905.