ChatGPT

Chiến lược Phân tích Dữ liệu Khám phá (EDA) cho Phân tích Tín dụng Ngân hàng

Giới thiệu

Phân tích dữ liệu khám phá (Exploratory Data Analysis – EDA) là bước đầu tiên và then chốt trong quy trình phân tích rủi ro tín dụng của ngân hàng. Mục tiêu của EDA là hiểu rõ cấu trúc và đặc điểm của dữ liệu khách hàng vay, từ đó phát hiện các mô hình, xu hướng và **dấu hiệu bất thường** trong hành vi tín dụng. Kết quả EDA giúp các tổ chức tài chính **đánh giá rủi ro khoản vay và khách hàng**, xây dựng hệ thống **chấm điểm tín dụng** và hỗ trợ quyết định **phê duyệt hoặc từ chối** khoản vay một cách khách quan. Theo một hướng dẫn về mô hình rủi ro tín dụng, các nhà phân tích nên bắt đầu bằng việc trực quan hóa dữ liệu với **biểu đồ hộp (boxplot)**, **biểu đồ histogram** và các thống kê mô tả để xác định **phân phối**, độ lệch (skewness) và các **điểm bất thường (outliers)** trong tập dữ liệu [1]. Việc làm sạch và thấu hiểu dữ liệu ngay từ đầu giúp đảm bảo tính chính xác của mô hình và tuân thủ các yêu cầu quản lý trong ngành tài chính.

Trong bối cảnh tín dụng thực tế, chiến lược EDA cần bám sát thông lệ của các hệ thống chấm điểm uy tín như **FICO** (Mỹ) hoặc **CIC** (Trung tâm Thông tin Tín dụng Quốc gia Việt Nam). Các hệ thống này dựa trên nhiều yếu tố về lịch sử tín dụng và nhân khẩu để đánh giá mức độ rủi ro của khách hàng. Ví dụ, điểm FICO – được 90% các định chế tài chính Hoa Kỳ sử dụng – tính toán dựa trên **5 nhóm thông tin chính**: lịch sử thanh toán (35%), số nợ đang có (30%), độ dài lịch sử tín dụng (15%), loại hình tín dụng (10%) và khoản tín dụng mới mở (10%) [2]. Những yếu tố này định hướng cho việc lựa chọn đặc điểm phân tích trong EDA. Tương tự, tại Việt Nam, CIC sử dụng mô hình chấm điểm nội bộ (403–706 điểm) để xếp hạng độ uy tín tín dụng của khách hàng thành các nhóm từ "Rất tốt" đến "Xấu" nhằm hỗ trợ các ngân hàng đánh giá rủi ro khi cho vay [3] [4].

Dưới đây, chúng ta đề xuất một chiến lược EDA chuẩn và chi tiết cho phân tích tín dụng, bao gồm việc phân tích theo **nhóm đặc điểm chính**, đề xuất các **biểu đồ trực quan** và phương pháp phân tích **định tính vs. định lượng**, nhận diện **tín hiệu cảnh báo sớm**, phân nhóm khách hàng theo **mức độ rủi ro (risk buckets)**, và gợi ý cách **xây dựng scorecard** hoặc kỹ thuật tạo đặc trưng (feature engineering) cho mô hình điểm tín dụng sau này.

Phân tích theo nhóm đặc điểm chính

Để có cái nhìn toàn diện, dữ liệu tín dụng nên được phân tích theo từng nhóm đặc điểm có liên quan đến **khả năng trả nợ** của khách hàng. Các nhóm chính thường gồm: (1) **Đặc điểm nhân khẩu học**, (2) **Đặc điểm tài chính cá nhân**, (3) **Hành vi tín dụng**, (4) **Lịch sử nợ xấu**, và (5) **Thông tin các khoản vay trước**. Việc tách bạch các nhóm này giúp xác định rõ vai trò của từng loại thông tin đối với rủi ro tín dụng.

Đặc điểm nhân khẩu học (demographics)

Nhóm đặc điểm nhân khẩu học bao gồm các yếu tố như **tuổi**, **giới tính**, **tình trạng hôn nhân**, **số người phụ thuộc**, **khu vực sinh sống**, v.v. Mặc dù ở một số thị trường (như Mỹ) các yếu tố như giới

tính, tình trạng hôn nhân không được sử dụng trực tiếp trong chấm điểm tín dụng do quy định chống phân biệt, dữ liệu nhân khẩu học vẫn hữu ích để hiểu bối cảnh khách hàng:

- **Tuổi của người vay**: Phân tích phân phối tuổi và so sánh giữa nhóm khách hàng trả nợ tốt và nhóm vỡ nợ. Thường có thể dùng **biểu đồ histogram** phân bố tuổi, kèm theo **biểu đồ hộp** để so sánh độ tuổi trung vị của hai nhóm (khách hàng tốt vs. xấu). Ví dụ, khách hàng quá trẻ có thể chưa có lịch sử tín dụng đủ dài, trong khi khách hàng lớn tuổi có thể sắp nghỉ hưu – cả hai thái cực tuổi tác đều có thể ảnh hưởng đến rủi ro tín dụng. Cần kiểm tra xem **tỷ lệ default** (vỡ nợ) có xu hướng cao hơn ở nhóm tuổi nào. Nếu dữ liệu cho thấy một ngưỡng tuổi nhất định có tỷ lệ vỡ nợ tăng đột biến, đó có thể là **tín hiệu cảnh báo** cần chú ý.

- **Giới tính**: Nếu được phép phân tích, có thể xem xét tỷ lệ nợ quá hạn giữa nam và nữ. Tuy nhiên, cần thận trọng vì yếu tố này có thể bị ảnh hưởng bởi cấu trúc mẫu. Thường thì phân tích này mang tính định tính, để hiểu đặc điểm khách hàng hơn là đưa trực tiếp vào mô hình.

- **Tình trạng hôn nhân và số người phụ thuộc**: Nhóm khách hàng đã kết hôn hoặc độc thân có hành vi tín dụng khác nhau? Ví dụ, người có gia đình có thể có trách nhiệm tài chính cao hơn (nuôi con, người phụ thuộc) và điều này có thể ảnh hưởng đến khả năng trả nợ. **Biểu đồ cột (bar chart)** có thể dùng để so sánh tỷ lệ default giữa các nhóm (độc thân, đã kết hôn, ly hôn...). Phân tích định lượng có thể bao gồm kiểm định Chi-square để xem sự khác biệt về tỷ lệ default giữa các nhóm này có ý nghĩa thống kê hay không.

- **Địa lý/khu vực**: Đôi khi dữ liệu có thông tin vùng miền hoặc thành thị/nông thôn. Kinh nghiệm cho thấy khu vực địa lý liên quan đến mức độ tiếp cận tín dụng và hành vi trả nợ. Ta có thể vẽ bản đồ nhiệt (heatmap) hoặc biểu đồ cột theo tỉnh/thành phố để xem **tỷ lệ nợ xấu** phân bố ra sao theo khu vực. Nếu một số khu vực có tỷ lệ nợ xấu cao bất thường, cần tìm hiểu thêm (có thể do kinh tế địa phương khó khăn hoặc chính sách tín dụng địa phương khác biệt).

Nhìn chung, phân tích nhóm nhân khẩu học chủ yếu mang tính **mô tả định tính**, giúp **phác họa chân dung khách hàng**. Tuy nhiên, cũng có thể lượng hóa một số phát hiện – chẳng hạn tính **hệ số tương quan** (phi-correlation cho biến phân loại) giữa tuổi và biến mục tiêu (default), hoặc xây dựng các **nhóm tuổi** (binned ages) để xem xu hướng default có **đơn điệu** theo tuổi hay không.

Đặc điểm tài chính cá nhân

Nhóm đặc điểm tài chính tập trung vào khả năng thu nhập và nghĩa vụ tài chính của khách hàng, bao gồm **thu nhập**, **tài sản** và các **tỷ số tài chính quan trọng**:

- **Thu nhập hàng tháng/năm**: Đây là biến liên tục cần được kiểm tra phân phối. Thường thu nhập có phân phối lệch phải (skewed) do một số ít khách hàng thu nhập rất cao. Ta nên vẽ **biểu đồ histogram** (có thể lấy log thu nhập để giảm độ lệch) để hiểu mặt bằng thu nhập. Tiếp đó, dùng **biểu đồ hộp** so sánh phân phối thu nhập của nhóm khách trả nợ tốt vs. nhóm vỡ nợ. Nếu nhóm vỡ nợ có thu nhập trung vị thấp hơn hẳn, điều này phù hợp với kỳ vọng rằng thu nhập thấp có thể khó trả nợ. Tuy nhiên cũng có trường hợp thu nhập cao nhưng vẫn vỡ nợ do chi tiêu quá mức – vì vậy cần xem xét thêm các chỉ số khác kết hợp.

- **Chi phí sinh hoạt ước tính và dư địa tài chính**: Nếu dữ liệu có thông tin về chi phí sinh hoạt hoặc dư nợ thẻ tín dụng hàng tháng, có thể tính **tỷ lệ thu nhập còn lại** sau khi trừ chi phí và nghĩa vụ nợ. Chỉ số này giúp đánh giá khách hàng còn bao nhiêu khả năng trả nợ. EDA nên xem

xét phân phối của chỉ số này và tìm ngưỡng nguy hiểm (ví dụ dưới 20% thu nhập còn lại có rủi ro cao).

- **Tỷ lệ Nợ trên Thu nhập (Debt-to-Income, DTI)**: Đây là **chỉ báo tài chính quan trọng** thường được sử dụng làm tiêu chí sàng lọc tín dụng sớm. DTI = Tổng nợ phải trả hàng tháng / Thu nhập hàng tháng. **Tỷ lệ DTI cao** cho thấy khách hàng gánh nhiều nợ so với thu nhập, là dấu hiệu cảnh báo về khả năng trả nợ trong tương lai [5]. Chẳng hạn, nếu DTI > 50% có thể xem là rất rủi ro. Ta nên vẽ **biểu đồ cột** cho các khoảng DTI (0-20%, 20-40%, 40-60%, >60%) và quan sát **tỷ lệ default** tăng theo các bucket DTI này. Một xu hướng tăng rõ rệt của tỷ lệ vỡ nợ ở nhóm DTI cao là một kết quả định lượng quan trọng. Đây cũng chính là một **tín hiệu cảnh báo sớm**: nếu **DTI tăng dần theo thời gian** với một khách hàng, họ có thể đang vay mượn quá mức [5].

- **Tài sản đảm bảo (Collateral)**: Đối với các khoản vay thế chấp (như vay mua nhà, vay mua ô tô), thông tin về tài sản đảm bảo và tỷ lệ giữa khoản vay trên giá trị tài sản (Loan-to-Value, LTV) cần được phân tích. **Collateral thấp** (giá trị tài sản thấp so với khoản vay) hoặc **LTV cao** (vay gần bằng hoặc vượt giá trị tài sản) là tín hiệu rủi ro vì nếu khách hàng default, tổ chức tín dụng khó thu hồi đủ nợ từ việc xử lý tài sản [5]. EDA nên kiểm tra phân phối LTV và tỷ lệ default tương ứng. Ví dụ, vẽ biểu đồ tỷ lệ default theo các khoảng LTV (dưới 50%, 50-80%, 80-100%, >100%). Nếu LTV vượt 100% (nợ nhiều hơn tài sản) chắc chắn nằm trong nhóm rủi ro cao.

- **Biểu đồ và phân tích bổ sung**: Sử dụng **ma trận tương quan** cho các biến tài chính (thu nhập, DTI, LTV, tài sản...) để xem chúng có liên hệ mạnh với nhau không (ví dụ thu nhập cao có khuynh hướng DTI thấp). Nếu có hiện tượng đa cộng tuyến cao giữa các biến tài chính, ta cần lưu ý khi đưa vào mô hình sau này (có thể phải chọn lọc hoặc biến đổi biến). Về mặt định lượng, có thể chạy **phân tích hồi quy tuyến tính** nhỏ giữa các biến tài chính với nhau để xác định quan hệ, nhưng đa phần EDA dừng ở mức tương quan và quan sát đồ thị.

Hành vi tín dụng (credit behavior)

Đây là nhóm đặc điểm rất quan trọng, phản ánh **cách thức khách hàng sử dụng và quản lý tín dụng**. Nhiều biến trong nhóm này có thể thu thập từ **báo cáo tín dụng** (credit bureau) hoặc lịch sử quan hệ tín dụng của khách hàng với ngân hàng:

- **Lịch sử thanh toán (Payment History)**: Đây thường là **yếu tố quan trọng nhất** trong chấm điểm tín dụng. Ví dụ, hệ thống FICO coi lịch sử thanh toán chiếm khoảng 35% trọng số điểm [2]. Cần phân tích các biến như: số lần **trả trễ hạn** (30 ngày, 60 ngày, 90 ngày), có từng bị **vỡ nợ/ charge-off** chưa, số tháng kể từ lần trễ hạn gần nhất, v.v. **Biểu đồ cột** có thể cho thấy tỷ lệ khách hàng default tăng mạnh nếu từng có lịch sử trễ hạn trước đó. Chẳng hạn, vẽ biểu đồ tỷ lệ vỡ nợ của khách hàng theo số lần trễ hạn (0 lần, 1 lần, >1 lần): ta kỳ vọng nhóm **chưa từng trả trễ** có tỷ lệ default thấp nhất, trong khi nhóm **nhiều lần trễ hạn** có tỷ lệ cao nhất – một **tín hiệu cảnh báo sớm** rõ ràng. Việc **phân nhóm nhị phân** (có/không từng trễ hạn) cũng hữu ích: khách hàng **có lịch sử nợ xấu** trước đây sẽ được đưa vào nhóm rủi ro cao hơn khi ra quyết định cấp tín dụng [5].

- **Mức dư nợ và sử dụng tín dụng (Credit Utilization)**: Mức dư nợ hiện tại trên các thẻ tín dụng và khoản vay của khách hàng so với hạn mức tín dụng của họ. **Tỷ lệ sử dụng tín dụng cao** (ví dụ dùng > 80% hạn mức thẻ) thường báo hiệu khách hàng đang **căng thẳng tài chính**. Phân tích EDA nên tính toán tỷ lệ sử dụng cho mỗi khách hàng và vẽ **biểu đồ hộp** cho tỷ lệ này ở hai nhóm outcome (default vs. không default). Nghiên cứu đã chỉ ra rằng nếu một người sử dụng hầu hết hạn mức tín dụng của mình, các ngân hàng có thể coi đó là dấu hiệu người đó đang vay mượn quá mức và rủi ro vỡ nợ cao hơn [6]. Đây chính là lý do FICO dành ~30% trọng số cho nhóm

"Amounts Owed" (số nợ đang có) [2]. Ta có thể trực quan hóa bằng **đồ thị phân tán** (scatter plot) giữa tỷ lệ sử dụng và xác suất default, hoặc đơn giản hơn là bảng tỷ lệ default theo các bucket sử dụng (0-30%, 30-60%, >60%). Kỳ vọng xu hướng: sử dụng càng cao thì tỷ lệ default càng lớn.

- **Số lượng tài khoản tín dụng**: Bao gồm số thẻ tín dụng mở, số khoản vay đã có, số dư nợ hiện tại. Khách hàng có quá nhiều khoản tín dụng mở có thể gặp khó khăn quản lý trả nợ. Tuy nhiên, có nhiều tài khoản cũng có thể đồng nghĩa khách hàng có kinh nghiệm tín dụng dày dặn. Do đó cần phân tích kỹ: vẽ histogram về số lượng tài khoản, và xem nhóm default có xu hướng có nhiều hay ít tài khoản hơn. Một cách khác là **phân tích độ dài lịch sử tín dụng** – tức là xem **số năm từ khi mở tài khoản tín dụng đầu tiên**. Lịch sử tín dụng dài (nhiều năm) thường giúp **giảm** rủi ro do khách hàng có kinh nghiệm (FICO coi yếu tố này ~15% trọng số điểm [2]). Ta có thể kiểm chứng điều này bằng cách so sánh **tuổi tín dụng trung bình** của hai nhóm. Nếu nhóm default có tuổi tín dụng ngắn hơn rõ rệt, điều đó phù hợp với giả thuyết.

- **Khoản tín dụng mới (New Credit)**: Số lượng tài khoản mới mở gần đây hoặc số lần bị tra cứu tín dụng (credit inquiries) trong thời gian ngắn. Khách hàng mở nhiều thẻ hoặc vay nhiều nơi trong vòng 3-6 tháng thường bị đánh giá rủi ro cao hơn (có thể đang rất cần tiền hoặc sắp quá khả năng vay). **Biểu đồ cột** cho thấy số lượng inquiry trong 6 tháng vs. tỷ lệ default sẽ hữu ích. Đây cũng chính là yếu tố "new credit" (~10% điểm FICO [2]). EDA có thể phát hiện ví dụ: khách hàng có >3 lần inquiry trong 6 tháng có tỷ lệ default cao gấp đôi bình thường – đó là insight quan trọng cho chính sách phê duyệt.

- **Cơ cấu loại tín dụng (Credit Mix)**: Loại hình khoản vay mà khách hàng đang có – ví dụ thẻ tín dụng, vay tiêu dùng, vay thế chấp, vay mua xe... Một **danh mục tín dụng đa dạng** (credit mix tốt) thường được đánh giá cao (điểm FICO cũng có phần dành cho credit mix 10% [2]). EDA có thể xem xét: nhóm khách hàng chỉ có vay tín chấp ngắn hạn có tỷ lệ default khác nhóm có vay thế chấp dài hạn không? **Biểu đồ chồng cột** (stacked bar) có thể biểu diễn tỷ trọng các loại khoản vay giữa hai nhóm khách hàng. Phân tích này mang tính định tính, giúp hiểu hành vi sử dụng tín dụng.

Nhìn chung, **hành vi tín dụng** là nhóm biến quan trọng nhất để nhận diện rủi ro. Phân tích định tính sẽ tập trung tìm các **xu hướng đáng ngờ** (ví dụ sử dụng hạn mức tối đa, mở nhiều khoản vay mới) còn định lượng sẽ hướng đến việc **định lượng sức phân biệt** của các biến này. Một kỹ thuật thường dùng trong ngành tài chính là tính **giá trị thông tin (Information Value – IV)** cho từng biến để xem biến nào dự báo tốt nhất khả năng default [7]. IV cao nghĩa là biến đó phân biệt rõ khách hàng tốt/xấu. Chẳng hạn, biến "số lần trễ hạn > 30 ngày" có thể có IV rất cao vì gần như mọi người default đều từng trễ hạn, trong khi nhóm tốt thì không. Chúng ta có thể tính IV trong EDA và dùng nó để chọn lọc biến đưa vào mô hình điểm sau này [7].

Lịch sử nợ xấu (các khoản nợ quá hạn/xóa nợ trước đây)

Nhóm này thực chất là một phần của hành vi tín dụng, nhưng cần tách riêng do tầm quan trọng đặc biệt: **nợ xấu** (thường hiểu là các khoản vay trước đây bị quá hạn hoặc không trả được). Dữ liệu từ CIC hoặc lịch sử nội bộ sẽ cho biết khách hàng từng có khoản vay nào bị quá hạn nghiêm trọng (nhóm nợ 3-5 theo CIC) hay không. Phân tích cần làm:

- **Tình trạng nợ xấu trong quá khứ**: Tạo một biến cờ (flag) đánh dấu khách hàng **từng có nợ xấu (có lịch sử xấu)** vs. **chưa từng**. Rõ ràng nhóm từng nợ xấu sẽ có rủi ro tái phạm cao hơn nhiều. **Biểu đồ cột** đơn giản so sánh tỷ lệ default hiện tại giữa hai nhóm sẽ cho thấy sự khác biệt. Đây là

một **tín hiệu cảnh báo sớm mạnh** – nếu hồ sơ CIC cho thấy khách hàng thuộc **nhóm nợ xấu** trước đây, hồ sơ đó gần như chắc chắn bị xếp vào nhóm rủi ro cao khi xét duyệt.

- **Mức độ nghiêm trọng và thời gian của nợ xấu**: Nếu có chi tiết, cần xem **mức độ nợ xấu cao nhất** từng gặp (nhóm 3: quá hạn 90 ngày, nhóm 4: cơ cấu nợ, nhóm 5: nợ mất vốn) và **thời gian đã trôi qua** từ lần đó. Khách hàng mới xoá nợ xấu gần đây rủi ro hơn khách hàng từng có nợ xấu 5 năm trước và đã phục hồi uy tín phần nào. EDA có thể dùng **bảng chéo** (cross-tab) giữa "có nợ xấu < 1 năm", "1-3 năm", ">3 năm" với outcome để xem thời gian có làm giảm rủi ro hay không.

- **Số lượng khoản nợ xấu**: Nếu khách hàng có nhiều hơn một khoản nợ xấu trong lịch sử, đó là điểm trừ rất lớn. Một lần nữa, biểu đồ cột phân bố số lượng nợ xấu (0, 1, >1) vs. outcome sẽ làm rõ điều này. Kết quả định lượng mong đợi: nếu có >1 nợ xấu, xác suất default hiện tại rất cao.

Lưu ý khi xử lý dữ liệu nợ xấu: cần kết hợp với **thông tin xếp hạng tín dụng bên ngoài** nếu có. Ví dụ, CIC đã có sẵn **điểm tín dụng và xếp hạng** của mỗi cá nhân, phản ánh lịch sử tín dụng tổng hợp. **Điểm tín dụng thấp** theo CIC (ví dụ dưới 550 điểm – thuộc nhóm "Dưới trung bình" hoặc "Xấu") là dấu hiệu khách hàng đã có lịch sử tín dụng không tốt [3] [4]. Ta có thể tích hợp điểm này vào EDA: xem phân phối điểm CIC trong dataset, và quan trọng hơn là tỷ lệ default của các nhóm điểm (rất tốt, tốt, trung bình, xấu). Nếu dữ liệu khớp logic, nhóm "Xấu" (điểm thấp nhất) sẽ trùng với nhóm default cao nhất.

Thông tin các khoản vay trước (lịch sử quan hệ tín dụng trước đây)

Nhóm đặc điểm này tập trung vào những **giao dịch vay mượn trước đây của khách hàng**, đặc biệt là trong quan hệ với **ngân hàng hiện tại** (nếu là khách hàng cũ) hoặc thông tin từ các nguồn khác:

- **Số lượng khoản vay đã từng vay**: Khách hàng đã có nhiều kinh nghiệm vay (đã vay và tất toán nhiều khoản trước đây) có thể phản ánh họ quen với việc trả nợ đúng hạn (nếu lịch sử tốt). Ngược lại, người chưa từng vay (tín dụng mới) có **rủi ro không chắc chắn** vì thiếu lịch sử để đánh giá. EDA: so sánh tỷ lệ default của **khách hàng mới (first-time borrower)** vs **khách hàng cũ**. Một bảng so sánh đơn giản sẽ hữu ích. Ngoài ra, trong nhóm khách hàng cũ, có thể xem **tỷ lệ tái vay**: khách hàng từng tất toán khoản vay và quay lại vay tiếp thường là tín hiệu tốt (vì nếu họ từng default, chắc ngân hàng không cho vay lại).

- **Hiệu suất khoản vay trước**: Nếu có dữ liệu chi tiết, ta nên xem **lịch sử trả nợ của các khoản vay trước** (tại ngân hàng hoặc qua CIC). Ví dụ: khách hàng A có 3 khoản vay trước, trong đó 1 khoản trả muộn 30 ngày. Điều này sẽ ảnh hưởng thế nào tới kết quả hiện tại? Phân tích có thể chi tiết: **tỷ lệ các khoản vay trước bị trễ hạn**. Nếu một người có 50% khoản vay trước từng trễ hạn, khả năng cao họ tiếp tục trễ hạn trong khoản vay mới.

- **Loại khoản vay trước**: Xem khách hàng đã quen với loại tín dụng nào. Nếu họ chỉ từng vay trả góp nhỏ, lần đầu vay khoản lớn (nhà đất) có thể rủi ro cao hơn. Đây là phân tích định tính nhiều hơn – xem hồ sơ khách hàng có phù hợp sản phẩm vay đang đề nghị không.

- **Thời gian gắn bó với ngân hàng**: Khách hàng gắn bó lâu năm, có nhiều sản phẩm (tiền gửi, thẻ, vay) thường có **mức độ uy tín** nhất định. EDA có thể cho thấy khách hàng có tài khoản lâu năm ít default hơn. Điều này cũng giống khía cạnh "lịch sử quan hệ" – tương tự điểm **Length of Credit History** đã nêu ở hành vi tín dụng.

- **Vintage Analysis (Phân tích theo lô khoản vay)**: Một kỹ thuật hữu ích để tìm **xu hướng ẩn** là phân tích theo **kỳ giải ngân (vintage)**. Ta có thể nhóm các khoản vay theo quý/năm giải ngân và theo dõi **tỷ lệ vỡ nợ sau X tháng**. Nếu EDA cho thấy các khoản vay giải ngân trong năm gần đây có tỷ lệ vỡ nợ nhanh hơn các năm trước (ví dụ sau 6 tháng, default rate 5% so với trước chỉ 3%), điều đó cảnh báo chất lượng xét duyệt đang giảm hoặc điều kiện thị trường xấu đi. Phân tích vintage thường trình bày bằng **đường cong tỷ lệ quá hạn theo thời gian** cho từng lứa khoản vay. Đây là cách phân tích **định lượng theo chuỗi thời gian**, giúp phát hiện sớm rủi ro danh mục.

Phương pháp trực quan hóa và phân tích định tính vs. định lượng

Để thực hiện các phân tích trên, cần lựa chọn **biểu đồ** và **phép phân tích thống kê** phù hợp. Dưới đây là một số phương pháp khuyến nghị:

- **Biểu đồ histogram & mật độ**: Dùng cho **biến liên tục** như tuổi, thu nhập, dư nợ... Nhìn histogram giúp hiểu được phân phối (chuông, lệch, đa đỉnh?). Chồng thêm biểu đồ mật độ của hai nhóm (good vs. bad) để so sánh trực quan sự khác biệt. Ví dụ, mật độ thu nhập của nhóm default có thể lệch về phía thấp hơn so với nhóm không default.

- **Biểu đồ hộp (Boxplot)**: Hữu ích để so sánh **phân phối và outlier** của biến liên tục giữa hai nhóm phân loại (default vs. không). Boxplot cho thấy trung vị, khoảng tứ phân vị và các giá trị ngoại lai. Nếu boxplot của "% sử dụng hạn mức" ở nhóm default nằm cao hơn hẳn nhóm không default, điều này **củng cố định lượng** cho giả thiết sử dụng cao dẫn đến rủi ro.

- **Biểu đồ cột và biểu đồ thanh (Bar/Column chart)**: Dùng cho **biến phân loại hoặc dữ liệu đã phân nhóm**. Ví dụ tỷ lệ default theo các hạng mục tình trạng hôn nhân, theo các nhóm điểm tín dụng CIC, hoặc số lần trễ hạn. Trục tung thường là tỷ lệ default (%), trục hoành là các danh mục. Biểu đồ cột rất trực quan để **phân tích định tính** (thấy ngay nhóm nào rủi ro cao) và có thể thêm **nhãn số liệu** để cung cấp thông tin định lượng.

- **Ma trận tương quan và heatmap**: Tính tương quan Pearson (cho biến liên tục) hoặc Cramer's V (cho biến phân loại) giữa các biến độc lập và cũng như với biến mục tiêu. Biểu đồ heatmap tương quan giúp **định tính** nhận ra nhóm biến nào tương tự nhau (có thể loại bớt biến trùng lắp). Đối với biến mục tiêu nhị phân, có thể tính **Point-Biserial correlation** hoặc dùng **IV** như đã đề cập để đánh giá sức mạnh phân biệt của từng biến. Bản đồ nhiệt IV cho 10 biến mạnh nhất sẽ cho cái nhìn định lượng về biến nào đáng chú ý.

- **Phân tích phân phối Good/Bad**: Trong tín dụng, người ta thường vẽ biểu đồ chồng cho thấy phân phối của biến X đối với nhóm good và bad. Ví dụ, vẽ hai đường mật độ cho biến "điểm bureau" – ta kỳ vọng đường của nhóm bad lệch về điểm thấp, nhóm good lệch về điểm cao. Ngoài ra có thể dùng **KS-statistic (Kolmogorov-Smirnov)** để định lượng mức độ tách biệt: KS càng cao (tối đa 100) nghĩa là hai phân phối càng khác nhau. Trong EDA, nếu một biến có KS giữa hai nhóm > 30, đó thường là biến phân biệt tốt.

- **Trực quan hóa phân khúc (segmentation)**: Kết hợp nhiều đặc điểm để tìm phân khúc rủi ro. Ví dụ, vẽ **bubble chart** với trục X là thu nhập, trục Y là DTI, kích thước bong bóng biểu thị điểm tín dụng CIC, màu sắc là kết quả (default hoặc không). Biểu đồ này cho **cái nhìn đa chiều** định tính – có thể thấy cụm khách hàng default tập trung ở vùng thu nhập thấp, DTI cao, điểm tín dụng thấp. Từ đó gợi ý xây dựng phân khúc rủi ro.

- **Phân tích định lượng nâng cao**: Khi cần, có thể dùng các kiểm định thống kê: kiểm định t-test/ Mann-Whitney U (so sánh trung bình hai nhóm), kiểm định Chi-square (xem biến phân loại và default có độc lập không), hoặc thậm chí thử chạy một mô hình hồi quy đơn biến để xem mức độ ảnh hưởng. Tuy nhiên, trong EDA, trọng tâm vẫn là **hiểu dữ liệu** chứ chưa phải **suy luận thống kê chính thức**, nên các kiểm định chỉ hỗ trợ thêm. Điều quan trọng là mọi kết luận EDA đều mang tính **thám hiểm**, cần được xác nhận lại khi xây dựng mô hình.

Một khía cạnh khác cần chú ý trong EDA là **xử lý dữ liệu thiếu và ngoại lệ**. Thống kê tỷ lệ thiếu dữ liệu ở mỗi biến, quyết định ngưỡng loại biến nếu thiếu quá nhiều (ví dụ >30% giá trị missing có thể loại bỏ biến đó) [8]. Kiểm tra ngoại lệ bằng boxplot hoặc phương pháp IQR, quyết định loại bỏ hay winsorize (phóng đại) các giá trị ngoại lai [1]. Trong tín dụng, ngoại lệ đôi khi lại là **khách hàng rủi ro cao thật** (ví dụ mức vay rất lớn, cực kỳ bất thường) chứ không phải lỗi dữ liệu, nên cần thận trọng khi loại bỏ – phải xem chúng có phải là sai sót hay **thể hiện nhóm rủi ro đặc thù**.

Tóm lại, kết hợp linh hoạt giữa **phân tích trực quan định tính** (đồ thị, quan sát mẫu hình) và **phân tích định lượng** (tính toán tỷ lệ, tương quan, IV, thống kê) sẽ cho bức tranh đầy đủ. EDA tốt sẽ làm nổi bật những yếu tố nào đáng chú ý nhất cho rủi ro tín dụng, làm tiền đề cho việc nhận diện **các tín hiệu cảnh báo sớm** và xây dựng mô hình phù hợp.

Các tín hiệu cảnh báo sớm (Early Warning Signals)

Thông qua EDA, nhà phân tích có thể nhận ra một số **tín hiệu cảnh báo sớm** về nguy cơ vỡ nợ của khách hàng. Đây là các dấu hiệu trong dữ liệu hiện tại báo hiệu khách hàng có thể gặp khó khăn tài chính trong tương lai gần. Một số tín hiệu quan trọng trong phân tích tín dụng gồm:

- **Lịch sử thanh toán kém**: Như đã đề cập, nếu khách hàng có **tiền sử thanh toán trễ hạn hoặc vỡ nợ** trước đây, đó là cảnh báo mạnh nhất. **Thanh toán trễ thường xuyên** là chỉ dấu khách hàng đang gặp vấn đề dòng tiền hoặc thiếu ý thức trả nợ, dự báo khả năng vỡ nợ cao [5]. EDA sẽ cho thấy ví dụ: chỉ cần 1 lần **quá hạn >30 ngày** cũng làm xác suất default tăng đáng kể. Vì vậy, chỉ cần phát hiện lịch sử này (qua CIC hoặc dữ liệu nội bộ) là đã có căn cứ thận trọng.

- **Nợ quá nhiều so với thu nhập (DTI cao)**: Nếu tỷ lệ DTI của khách hàng liên tục tăng và vượt ngưỡng an toàn (thường ~40-50%), đây là **tín hiệu sớm** cho thấy họ đang vay mượn thêm và **sức chịu đựng tài chính** giảm [5]. EDA có thể phát hiện qua việc so sánh DTI hiện tại với DTI quá khứ (nếu dữ liệu theo thời gian) hoặc so với mức chung của nhóm. DTI cao đặc biệt nguy hiểm khi lãi suất thị trường tăng – chi phí trả nợ sẽ vượt quá khả năng.

- **Tỷ lệ sử dụng hạn mức tín dụng cao**: Khách hàng **đang sử dụng gần hết hạn mức thẻ tín dụng** hoặc hạn mức thấu chi là tín hiệu họ không có dư địa tài chính dự phòng. Như FICO phân tích, việc dùng quá nhiều hạn mức có thể khiến điểm tín dụng giảm và báo hiệu nguy cơ vỡ nợ cao [6]. Nếu EDA cho thấy khách hàng đang max-out thẻ tín dụng, đó là lá cờ đỏ.

- **Nhiều khoản vay mới và yêu cầu tín dụng dồn dập**: Trong một khoảng thời gian ngắn (vài tháng), nếu khách hàng mở nhiều khoản vay mới hoặc có nhiều lần bị truy vấn tín dụng (inquiry) từ các tổ chức cho vay, điều này thường là **dấu hiệu căng thẳng tài chính**. Họ có thể đang "đảo nợ" hoặc tìm nguồn tiền khẩn cấp, làm tăng xác suất vỡ nợ. EDA nên phát hiện số lượng inquiry bất thường hoặc số tài khoản tăng đột biến gần đây.

- **Điểm tín dụng (credit score) giảm hoặc ở mức thấp**: Điểm tín dụng tổng hợp (như điểm CIC hoặc FICO) là chỉ báo gọn về nhiều khía cạnh. **Điểm thấp** rõ ràng là tín hiệu rủi ro – ví dụ điểm

CIC dưới 545 thuộc nhóm "Dưới trung bình" hoặc tệ hơn [9]. Nếu EDA cho thấy phần lớn khách default có điểm thấp, ta nên sử dụng ngay điểm này như một cờ cảnh báo. Ngược lại, nếu có trường hợp điểm cao mà vẫn default, cần điều tra nguyên nhân (có thể do sự kiện bất thường như mất việc đột ngột không phản ánh trong điểm quá khứ).

- **Dòng tiền âm hoặc giảm (Cash flow xấu)**: Đối với khách hàng doanh nghiệp hoặc hộ kinh doanh, dấu hiệu sớm là **doanh thu hoặc lợi nhuận suy giảm** liên tục, hoặc dòng tiền thuần chuyển âm [10]. Với khách hàng cá nhân, không có báo cáo tài chính chi tiết, nhưng có thể suy luận qua việc thu nhập giảm (nếu cập nhật) hoặc tăng chi tiêu tín dụng. Nếu ngân hàng có dữ liệu giao dịch tài khoản thanh toán của khách, có thể phát hiện thu nhập ròng hàng tháng đang giảm – đây cũng là tín hiệu cần chú ý.

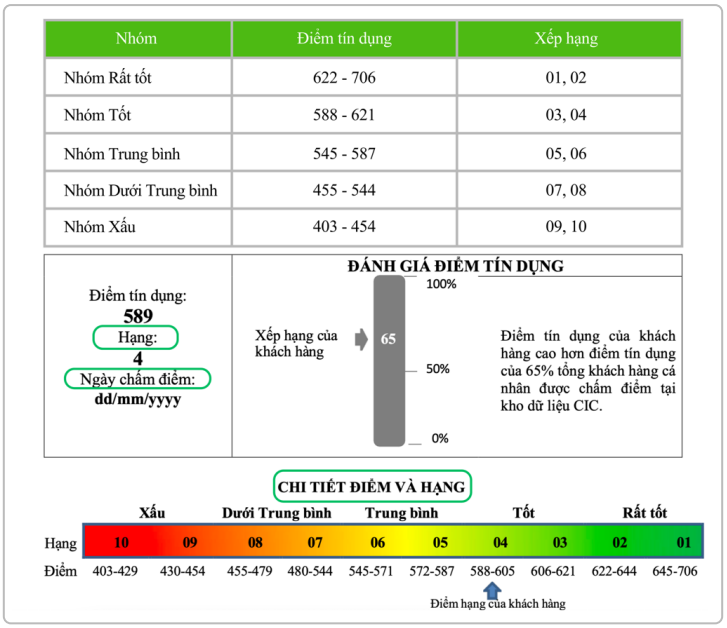
- **Thay đổi tình hình kinh tế, ngành nghề**: Mặc dù đây là yếu tố bên ngoài, EDA vĩ mô có thể cảnh báo sớm. Ví dụ, **tỷ lệ thất nghiệp tăng** hoặc **ngành nghề của khách hàng đang suy thoái** (như du lịch mùa COVID) sẽ làm tăng rủi ro vỡ nợ cho nhóm khách hàng tương ứng [10]. Trong phân tích, có thể gắn nhãn ngành nghề rủi ro cao (dựa trên thống kê vĩ mô) và xem nhóm khách đó có tỷ lệ default tăng không. Nếu có, khi xét duyệt nên khắt khe hơn với người trong ngành đó ở thời điểm hiện tại.

- **Thay đổi thông tin cá nhân bất lợi**: Với doanh nghiệp, thay đổi cơ cấu quản lý hoặc chủ sở hữu có thể báo hiệu rủi ro [11]. Với cá nhân, có thể không rõ ràng như vậy, nhưng các sự kiện như đổi công việc liên tục, chuyển địa chỉ thường trú, v.v. nếu ghi nhận được cũng có thể xem là dấu hiệu không ổn định.

Những tín hiệu cảnh báo sớm trên cần được **theo dõi liên tục**. Trong thực tế, các ngân hàng tiên tiến thiết lập **hệ thống Early Warning System (EWS)** để giám sát các chỉ số này hàng tháng/quý trên toàn bộ danh mục tín dụng [12] [10]. Tuy nhiên, ngay ở bước EDA, ta đã có thể xác định và **đánh dấu cờ đỏ** cho từng hồ sơ khách hàng nếu họ xuất hiện các đặc điểm rủi ro. Khi đưa vào hệ thống phê duyệt, các cờ đỏ này sẽ giúp chuyên viên tín dụng tập trung điều tra kỹ hơn trước khi quyết định.

Phân nhóm khách hàng theo mức độ rủi ro (Risk Buckets)

Một kết quả quan trọng của EDA trong phân tích tín dụng là khả năng **phân nhóm khách hàng theo mức độ rủi ro**, thường gọi là **risk buckets** hoặc **risk segments**. Việc phân nhóm này dựa trên các đặc điểm và tín hiệu rủi ro đã phân tích, giúp ngân hàng **ra quyết định đồng nhất** cho mỗi nhóm và **tinh chỉnh chiến lược** (lãi suất, hạn mức, quy trình phê duyệt) phù hợp với từng mức rủi ro.



Hình 1: Ví dụ phân loại khách hàng theo điểm tín dụng CIC thành 5 nhóm xếp hạng rủi ro, từ "Rất tốt" đến "Xấu" (tương ứng 10 hạng điểm chi tiết) [3] [4] *. Mỗi nhóm được định nghĩa bởi khoảng điểm tín dụng; điểm càng cao thì xếp hạng càng tốt, rủi ro vỡ nợ càng thấp.*

Trong thực tế, có nhiều cách phân chia risk bucket. **Cách phổ biến** là dựa vào **điểm số rủi ro** (risk score) của mỗi khách hàng. Điểm rủi ro có thể là **điểm tín dụng bên ngoài** (như FICO, CIC) hoặc **xác suất vỡ nợ (PD)** dự đoán từ mô hình nội bộ. Ví dụ, CIC chia điểm 403–706 thành 5 nhóm chính: Rất tốt, Tốt, Trung bình, Dưới trung bình, Xấu, kèm 10 hạng chi tiết (01 đến 10) [3] [4]. Tương tự, FICO 300–850 thường được các tổ chức chia thành các nhóm: <580 (kém), 580-669 (trung bình/yếu), 670-739 (khá), 740-799 (tốt), 800+ (xuất sắc). Mục đích là **đơn giản hóa việc ra quyết định**: ví dụ, với nhóm "tốt" trở lên thì tự động duyệt vay với lãi suất ưu đãi, nhóm "trung bình" thì duyệt có điều kiện (yêu cầu thêm tài sản đảm bảo chẳng hạn), còn nhóm "xấu" thì từ chối. Phân nhóm như vậy giúp xử lý số lượng lớn đơn vay một cách nhanh chóng và nhất quán.

Từ kết quả EDA, ta có thể đề xuất ngưỡng phân nhóm. Chẳng hạn, nếu nhận thấy **PD ước tính 1 năm** trên 10% là rủi ro cao, 5-10% trung bình, dưới 5% là thấp, ta có thể chia 3 nhóm tương ứng. Hoặc dựa trên **điểm số từ mô hình** (score 300-900 điểm), ta chia thành các bậc 300-500 (xấu), 501-600 (trung bình), 601-700 (tốt), >700 (rất tốt). Việc chọn số lượng nhóm và ranh giới cần dựa trên **phân phối thực tế** của điểm rủi ro và **mục tiêu kinh doanh** (ví dụ muốn duy trì tỷ lệ chấp nhận bao nhiêu phần trăm khách).

EDA cũng có thể khám phá **đặc điểm của từng phân khúc rủi ro**. Ví dụ, lấy phân nhóm theo điểm CIC ở trên: phân tích đặc điểm nhân khẩu, tài chính của nhóm "Xấu" xem có gì nổi bật (có thể thu nhập thấp, nhiều nợ, lịch sử tín dụng ngắn...). Từ đó, giúp **định hình chính sách** cho nhóm này: có thể cần sản phẩm tín dụng khác phù hợp hơn (như cho vay có bảo đảm hoặc đồng thời tư vấn tài chính cho họ). Ngược lại, nhóm "Rất tốt" có thể được ưu tiên tăng hạn mức, cross-selling thêm sản phẩm.

Một lưu ý là các **ngưỡng risk bucket** không nên quá cứng nhắc. Ngân hàng có thể áp dụng thêm các **quy tắc chuyên gia** dựa trên kinh nghiệm. Ví dụ: dù điểm cao nhưng nếu phát hiện khách có **cờ đỏ** (như nợ xấu mới hồi phục) thì vẫn đẩy họ vào bucket rủi ro cao hơn để xem xét kỹ. EDA cung cấp dữ liệu nền tảng, còn việc phân bucket cuối cùng thường kết hợp cả **phân tích định lượng** và **kinh nghiệm định tính**.

Nhìn chung, phân nhóm rủi ro là cầu nối giữa EDA và thực thi quyết định. Việc này cũng hỗ trợ trong quản lý danh mục: ngân hàng có thể theo dõi **tổn thất kỳ vọng** từ mỗi bucket, phân bổ hạn mức tín dụng hợp lý và **định giá lãi suất theo rủi ro** (risk-based pricing). Các nghiên cứu cho thấy phương pháp logistic regression đưa ra xác suất default, và người ta thường dùng trực tiếp xác suất đó làm **điểm rủi ro** để phân loại khách hàng vào các risk bucket khác nhau [13]. Việc áp dụng thống nhất như vậy đảm bảo **tính khách quan và tuân thủ** trong quá trình phê duyệt tín dụng.

Đề xuất mô hình điểm tín dụng và Feature Engineering

Sau khi hoàn thành EDA và xác định được những đặc tính quan trọng, bước tiếp theo thường là xây dựng **mô hình điểm tín dụng (credit scoring model)**. Mô hình này sẽ gán cho mỗi khách hàng một điểm số rủi ro hoặc xác suất default dự đoán, dùng để ra quyết định cấp tín dụng. Dựa trên thông lệ quốc tế và yêu cầu quản lý, phương pháp **Logistic Regression (hồi quy logistic)** rất được ưa chuộng trong phát triển scorecard tín dụng [14]. Lý do chính là logistic regression vừa cho kết quả tương đối chính xác, vừa **dễ giải thích** – một yếu tố quan trọng khi ngân hàng phải giải trình quyết định từ chối cấp tín dụng cho khách hàng (yêu cầu đưa ra "lý do từ chối" rõ ràng) [14]. Cơ quan quản lý và các hệ thống như **Basel** cũng khuyến khích mô hình dễ diễn giải, tránh tình trạng "hộp đen". Vì vậy, dù hiện nay có nhiều kỹ thuật Machine Learning phức tạp, logistic regression kết hợp với scorecard vẫn là "tiêu chuẩn vàng" trong nhiều tổ chức tín dụng.

Feature Engineering (xây dựng đặc trưng) là bước cầu nối từ EDA sang mô hình. Một kỹ thuật phổ biến trong scorecard là chuyển các biến thô thành **điểm số con** thông qua **binning + Weight of Evidence (WOE)**. Cụ thể, mỗi biến (tuổi, thu nhập, DTI, v.v.) được chia thành các **nhóm giá trị rời rạc** (bins) sao cho tỷ lệ default trong các bin tăng dần đều (tính **đơn điệu**). Sau đó, mỗi bin được gán một **giá trị WOE** phản ánh mức độ rủi ro tương đối của bin đó. Mô hình logistic sẽ sử dụng các WOE này làm đầu vào. Việc này giúp đảm bảo quan hệ biến-đích là **đơn điệu và tuyến tính** trên log-odds, thỏa mãn giả định của logistic và tránh những **nghịch lý** khi giải thích mô hình [15]. Ví dụ, biến **DTI** có thể được bin: DTI < 30%, 30-50%, >50%. Nếu mô hình cho điểm ngược (DTI trung bình được duyệt nhưng DTI thấp lại bị từ chối) thì rõ ràng là bất hợp lý – kỹ thuật binning và ép đơn điệu sẽ ngăn chặn điều này [16] [17].

Một **scorecard framework** điển hình sẽ bao gồm: danh sách các đặc trưng đã chọn (sau EDA và tính IV, thường chọn ~10-20 biến tốt nhất), mỗi đặc trưng được chia bin có trật tự rủi ro tăng dần, bảng tra WOE và **điểm số** cho từng bin. Tổng điểm của khách hàng = tổng điểm từng đặc trưng. Thang điểm thường được hiệu chỉnh về dạng điểm quen thuộc (ví dụ 300-850 như FICO, hoặc 1-100). Trong quá trình feature engineering, cần lưu ý **loại bỏ hoặc hợp nhất** các biến có ý nghĩa tương tự (để tránh double-count rủi ro) và đảm bảo mô hình tuân thủ quy định (ví dụ không dùng biến nhạy cảm, và biến giải thích phải hợp lý về mặt kinh doanh).

Kết quả EDA có thể gợi ý một số **biến tương tác** hoặc **biến phái sinh** hữu ích. Chẳng hạn, nếu EDA cho thấy mối quan hệ giữa thu nhập và default phụ thuộc vào ngành nghề (ví dụ thu nhập thấp trong ngành xây dựng rủi ro hơn thu nhập thấp ngành công chức), ta có thể tạo biến kết hợp "thu nhập*ngành nghề". Tuy nhiên, các mô hình truyền thống thường ưu tiên sự đơn giản, nên chỉ tạo biến tương tác khi thực sự cần. Một biến phái sinh thông dụng từ EDA là tổng điểm hành vi*: ví dụ đếm số tín hiệu cảnh báo sớm mỗi khách hàng có (trễ hạn, DTI cao, sử dụng hạn mức cao...) – nếu >2 tín hiệu thì đánh cờ đỏ. Biến này có thể đưa vào mô hình như một đặc trưng tổng hợp rủi ro.

Trước khi hoàn tất mô hình, cần quay lại dữ liệu và chia tập **train/test** (hoặc train/validate/test) một cách hợp lý (thường dùng phân tầng theo thời gian để tránh rò rỉ tương lai [18]). Sau đó, **huấn luyện mô hình logistic** trên tập train, tinh chỉnh trên validate, và **đánh giá** trên tập test bằng các chỉ số: độ chính xác, AUC, Gini, KS. Mô hình tốt phải có tính phân biệt cao (KS, Gini cao) và giữ được tính ổn định khi

kiểm tra trên các khoảng thời gian khác nhau (tránh overfit). EDA hỗ trợ cả giai đoạn đánh giá này, ví dụ vẽ **ROC curve**, tính **gini coefficient** để so sánh với các mô hình khác [19].

Cuối cùng, mô hình được triển khai dưới dạng **scorecard** đơn giản để sử dụng tại hiện trường: nhân viên tín dụng nhập các thông tin của khách vào, hệ thống tính điểm và đưa ra **khuyến nghị quyết định** (duyệt/từ chối hoặc chuyển lên cấp trên xét). Mô hình logistic/scorecard với điểm số cũng dễ dàng phân loại thành **risk bucket** như đã nói: ví dụ điểm > 700 auto duyệt, 600-700 cần xem xét thủ công, <600 thì từ chối. Điều này nhất quán với cách các tổ chức như FICO vận hành – sử dụng **điểm số rủi ro để xếp hạng khách hàng** và quản lý danh mục một cách khoa học [13].

Kết luận

Chiến lược EDA trong phân tích tín dụng ngân hàng đòi hỏi cách tiếp cận **toàn diện và có hệ thống**, kết hợp hiểu biết nghiệp vụ tài chính với kỹ thuật phân tích dữ liệu. Bằng cách khám phá dữ liệu theo từng nhóm đặc điểm – từ nhân khẩu học, tài chính cá nhân đến hành vi tín dụng và lịch sử nợ xấu – chúng ta có thể xác định những yếu tố cốt lõi ảnh hưởng đến rủi ro tín dụng. Sử dụng các biểu đồ trực quan và phương pháp thống kê phù hợp giúp làm nổi bật sự khác biệt giữa khách hàng tốt và xấu, qua đó nhận diện các **tín hiệu cảnh báo sớm** như DTI cao, thanh toán trễ, sử dụng tín dụng tối đa hay điểm tín dụng thấp.

Những phát hiện từ EDA tạo nền tảng để phân nhóm khách hàng theo **mức độ rủi ro** một cách hợp lý, tương đồng với các hệ thống chấm điểm chuẩn như FICO hay CIC, giúp chuẩn hóa quyết định phê duyệt và quản trị danh mục. Hơn nữa, EDA định hướng cho việc thiết kế **mô hình điểm tín dụng** – từ khâu chọn biến, tạo đặc trưng (WOE/IV) đến hiệu chỉnh mô hình sao cho vừa **hiệu quả dự báo** vừa **tuân thủ quy định**.

Tài liệu và thông lệ ngành cho thấy, một quy trình EDA bài bản không chỉ cải thiện chất lượng mô hình tín dụng mà còn nâng cao sự hiểu biết về khách hàng, cho phép ngân hàng **chủ động hơn trong quản lý rủi ro**. Điều này đặc biệt quan trọng trong bối cảnh thị trường biến động, khi mà việc sớm nhận biết rủi ro và phân loại khách hàng chính xác sẽ giúp tổ chức tín dụng duy trì **nợ xấu ở mức kiểm soát**, đồng thời không bỏ lỡ cơ hội mở rộng tín dụng một cách an toàn [20] [21]. Tóm lại, chiến lược EDA tín dụng hiệu quả sẽ là bước đệm vững chắc tiến tới xây dựng hệ thống chấm điểm và quản trị rủi ro hiện đại, hỗ trợ ngân hàng **ra quyết định sáng suốt** và **phát triển bền vững** trong hoạt động cho vay.

Nguồn tài liệu tham khảo: Các thông tin và thông lệ nêu trên được tổng hợp từ hướng dẫn và nghiên cứu trong ngành tài chính – ngân hàng, bao gồm tài liệu giáo khoa về phát triển mô hình **scorecard**, khuyến nghị của các tổ chức tín dụng uy tín, cũng như ví dụ thực tế từ hệ thống điểm tín dụng **FICO (Hoa Kỳ)** và **CIC (Việt Nam)** [2] [3] [4]. Các nguyên tắc về phân tích biến số và cảnh báo rủi ro sớm được tham chiếu từ case study trong lĩnh vực ngân hàng [7] [5], cùng với kinh nghiệm thực tiễn được chia sẻ bởi các chuyên gia phân tích tín dụng. Chúng tôi cũng lưu ý tuân thủ các yêu cầu quản lý (như Basel, CIC) khi đề xuất chiến lược, nhằm đảm bảo tính khả thi và hiệu quả khi áp dụng vào hệ thống thực tế. [14] [15]

---

[1] [18] [19] Step-by-Step Guide to Credit Risk Modeling
https://www.mezzi.com/blog/step-by-step-guide-to-credit-risk-modeling

[2] [6] How are FICO Scores Calculated? | myFICO
https://www.myfico.com/credit-education/whats-in-your-credit-score

3 4 9 Hướng dẫn đọc báo cáo CIC - Phần 2: thông tin điểm tín dụng và xếp hạng | FE CREDIT - VAY TIÊU DÙNG TÍN CHẤP
https://www.fecredit.com.vn/tin-tuc-khuyen-mai/tin-tuc/huong-dan-doc-bao-cao-cic-phan-2-thong-tin-diem-tin-dung-va-xep-hang/

5 10 11 12 Early Warning Signal in Lending - Roopya - Data Driven Loan Origination & Underwriting Platform
https://roopya.money/early-warning-signal/

7 YOU CANalytics | Information Value (IV) & Weight of Evidence (WOE) - Banking Case Study
https://ucanalytics.com/blogs/information-value-and-weight-of-evidencebanking-case/

8 Credit Risk 1: Data Preparation and Exploratory Analysis | by Gawain Gan | Medium
https://medium.com/@gawaingan/credit-risk-1-data-cleaning-and-eda-85f8e583b5c2

13 14 15 16 17 Logistic Regression in the Credit Risk Industry | by 2nd Order Solutions | Medium
https://2os.medium.com/logistic-regression-in-the-credit-risk-industry-6eb27bc2784c

20 21 Taktile - Beginner’s guide to lending: How to assess credit risk
https://taktile.com/articles/beginners-guide-to-lending-how-to-assess-credit-risk