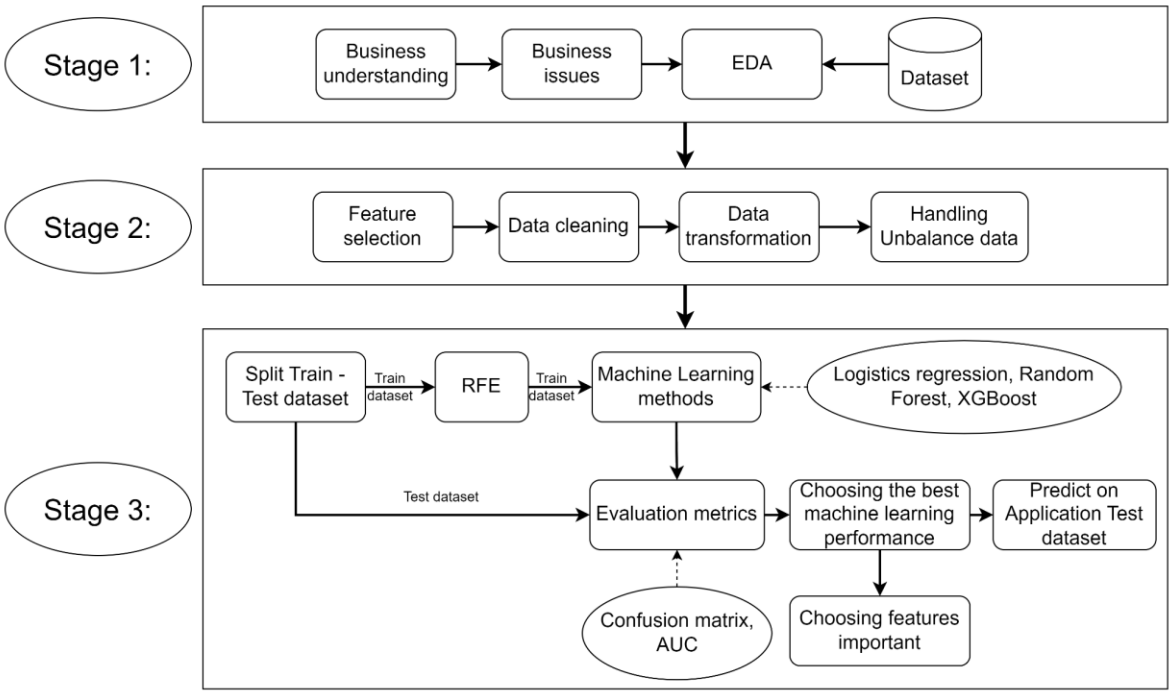
TÊN ĐỀ TÀI:
PHÂN TÍCH DỮ LIỆU TÀI CHÍNH VÀ ĐỀ XUẤT MÔ HÌNH DỰ ĐOÁN RỦI RO VÀ CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN VAY TÀI CHÍNH: TRƯỜNG HỢP CỦA CÔNG TY TÀI CHÍNH HOME CREDIT

CHƯƠNG 1: PHƯƠNG PHÁP THỰC HIỆN



Mô hình nghiên cứu

Dataset

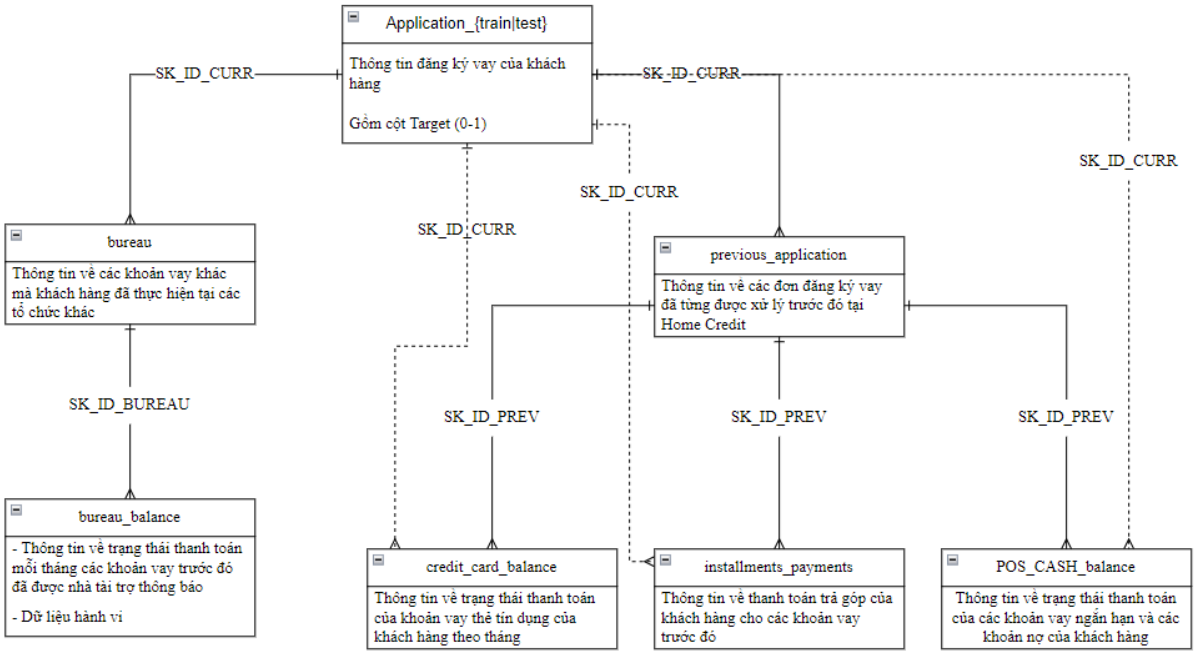
Home Credit Dataset, được thu thập từ Kaggle, với 8 bảng dữ liệu với nội dung từng bảng được mô tả chi tiết qua Bảng:

Tên bảng	Nội dung	Records	Thuộc tính
application_{train|test}.csv	Chứa thông tin về khách hàng đăng ký vay tiền.	307,511 quan sát trong tập train| 48,744	122 thuộc tính

Tên bảng	Nội dung	Records	Thuộc tính
		quan sát trong tập test	
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.	10,001,358 quan sát	1 thuộc tính phân loại và 7 biến giao dịch
bureau.csv	Chứa thông tin về các khoản vay khác mà khách hàng đã thực hiện tại các tổ chức khác.		17 thuộc tính
bureau_balance.csv	Chứa thông tin về trạng thái thanh toán các khoản vay trước đó đã được nhà tài trợ thông báo.		3 thuộc tính
credit_card_balance.csv	Chứa thông tin về trạng thái thanh toán của khoản vay thẻ tín dụng của khách hàng.	3,840,312 quan sát	23 thuộc tính giao dịch
installments_payments.csv	Chứa thông tin về thanh toán trả góp của khách hàng cho các khoản vay trước đó.	13,605,401 quan sát	8 thuộc tính giao dịch
previous_application.csv	Chứa thông tin về các đơn đăng ký vay đã từng được xử lý trước đó.	1,670,214 quan sát	37 thuộc tính giao dịch

Bảng 1.1. Các bảng trong dataset Home Credit

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



Sơ đồ mối quan hệ thực thể (ERD)

CHƯƠNG 2: PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA)

2.1. Bảng credit_card_balance

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.

2.2. Bảng POS_CASH_balance

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.

2.3. Bảng previous_application

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.

2.4. Bảng installment_payments

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.

2.5. Bảng application_train

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.

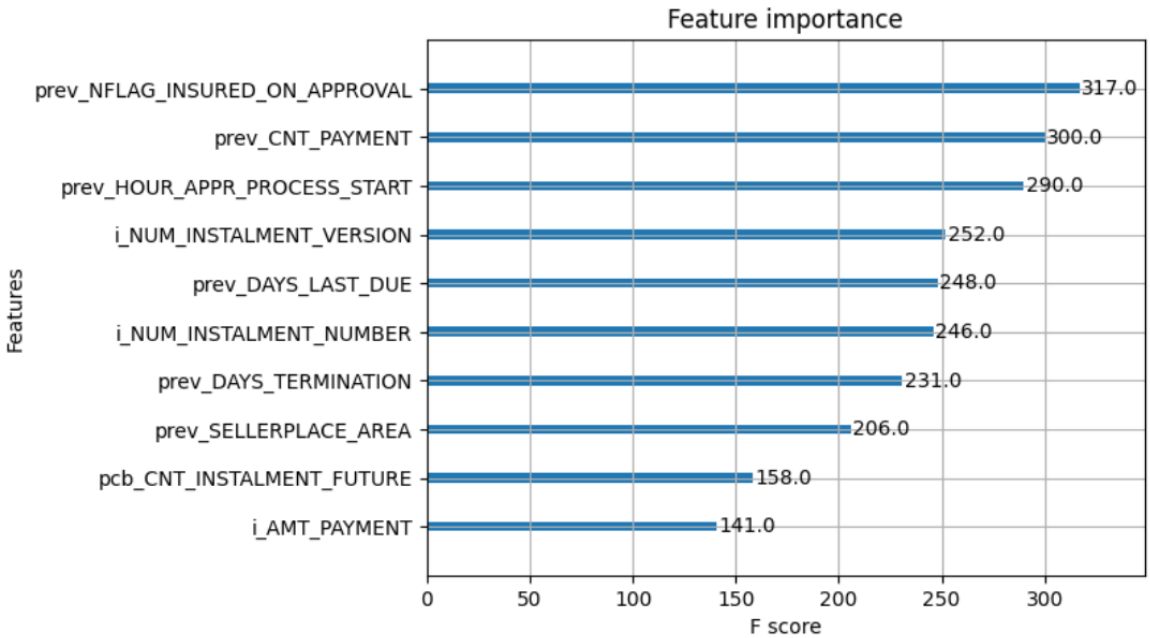
CHƯƠNG 3: KẾT QUẢ VÀ THẢO LUẬN

Đặc biệt, XGBoost đạt được hiệu suất cao với chỉ số Accuracy và F1-Score. 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 3.1. 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 3.1. 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.