DỰ BÁO RỦI RO VỠ NỢ CỦA KHÁCH HÀNG HOME CREDIT

Quá trình phân tích và xây dựng mô hình được thực hiện trong 3 bước bao gồm:

1. Xử lý dữ liệu
2. Phương Pháp Xây Dựng Mô Hình
3. Huấn luyện Mô Hình

- Chương trình thực hiện: **WoE_IV_Model.ipynb**
- Kết quả binning: **ReportVariable.xlsx**

1. Xử lý dữ liệu

Dữ liệu được thực hiện để phân tích và xây dựng mô hình được lấy từ Kaggle.com

- Bộ dữ liệu của có các đặc điểm sau:
- Số lượng quan sát: 307,511
- Số lượng Feature: 122
- Biến mục tiêu: label

Các biến của bộ dữ liệu thể hiện các thông tin bao gồm:

- Thông Tin Cá Nhân
- Thông Tin Tài Chính
- Thông Tin về Khoản Vay
- Thông Tin Liên Quan Đến Gia Đình
- Thông Tin Liên Quan Đến Việc Làm
- Thông Tin Về Lịch Sử Tín Dụng

Một số biến cần chú ý bao gồm:

Biến	Thông tin
SK_ID_CURR	ID duy nhất cho mỗi hồ sơ vay.
TARGET	Biến mục tiêu (1 có thể nghĩa là vỡ nợ, 0 nghĩa là không).
NAME_CONTRACT_TYPE	Loại hợp đồng khoản vay (ví dụ: vay tiền mặt, vay quay vòng)
WEEKDAY_APPR_PROCESS_START	Ngày trong tuần người vay nộp đơn.
CODE_GENDER	Giới tính của người vay.

FLAG_OWN_CAR	Chỉ ra liệu người vay có sở hữu xe hơi hay không
FLAG_OWN_REALTY:	Chỉ ra liệu người vay có sở hữu bất động sản hay không
CNT_CHILDREN	Số con của người vay
AMT_INCOME_TOTAL	Tổng thu nhập hàng năm của người vay
ORGANIZATION_TYPE	Loại hình tổ chức mà người vay làm việc
AMT_CREDIT	Tổng số tiền của khoản vay
AMT_ANNUITY	Số tiền trả góp hàng năm/kỳ hạn của khoản vay
AMT_GOODS_PRICE	Tổng giá trị của hàng hóa mà khoản vay nhằm mục đích mua
DAYS_BIRTH	Số ngày từ ngày sinh đến ngày làm hồ sơ vay (thường là giá trị âm)
DAYS_EMPLOYED	Số ngày từ ngày bắt đầu làm việc đến ngày làm hồ sơ vay (thường là giá trị âm).
NAME_HOUSING_TYPE	Loại hình nhà ở của người vay (ví dụ: nhà riêng, thuê nhà)
NAME_FAMILY_STATUS	Tình trạng hôn nhân của người vay
NAME_INCOME_TYPE:	Loại hình thu nhập của người vay (ví dụ: làm công ăn lương, tự kinh doanh)
NAME_TYPE_SUITE	Loại người đi cùng người vay khi làm hồ sơ vay (ví dụ: vợ/chồng, con cái)
NAME_EDUCATION_TYPE	Trình độ học vấn của người vay
OCCUPATION_TYPE	Loại nghề nghiệp của người vay

Xử lý Missing Data

Thực hiện xác định những biến bị missing data (xác định bằng % missing). Tiến hành xóa các trường dữ liệu lớn hơn 50%. Xác định được 41 biến bị missing nhiều hơn 50%.

2. Phương Pháp Xây Dựng Mô Hình

Phương pháp xây dựng mô hình áp dụng phương pháp hồi quy Logistic kết hợp với phân tích WoE. Phương pháp này không chỉ giúp dự đoán mối quan hệ giữa biến phụ thuộc nhị phân và biến độc lập mà còn tận dụng sức mạnh của phân tích WoE để đánh giá độ ảnh hưởng của từng biến đến mô hình.

Phân Tích WoE và Chỉ Số Information Value

- Sử dụng phân tích WoE để chuyển đổi giá trị của từng biến thành trọng số, thể hiện sức mạnh của biến đó đối với mục tiêu dự đoán.
- Chỉ số Information Value (IV) được tính toán để đánh giá chất lượng dự đoán của từng biến. Giá trị IV cung cấp thông tin về khả năng phân loại của biến và đóng vai trò quan trọng trong quá trình lựa chọn biến cho mô hình.

Lựa Chọn Biến Mạnh và Xác Định Ngưỡng IV

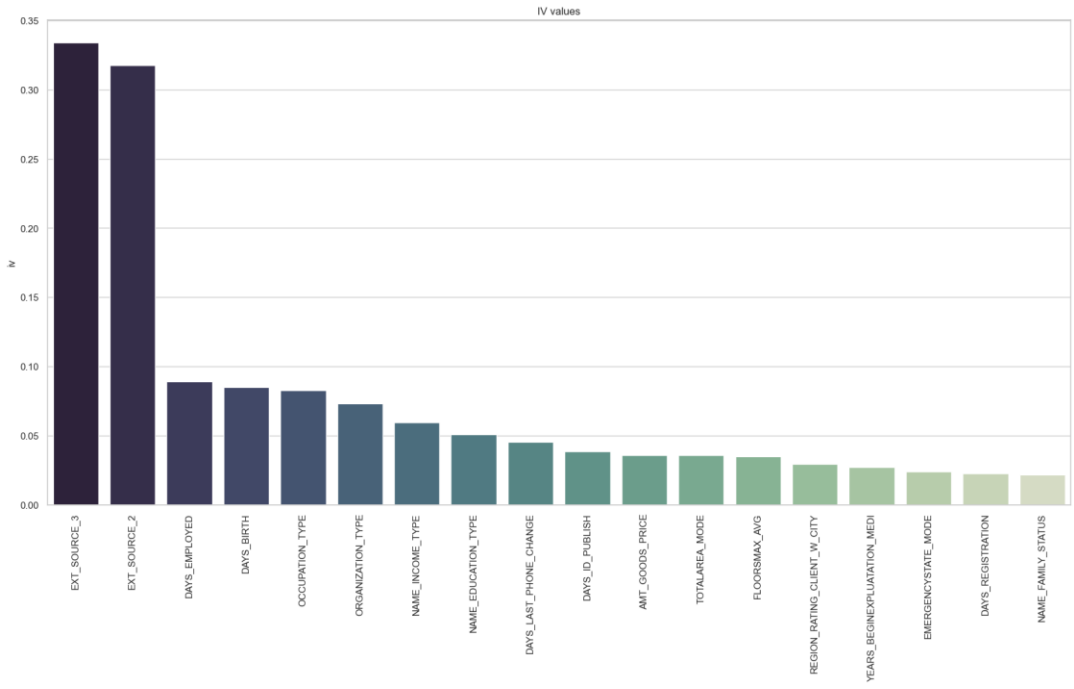
- Dựa vào kết quả phân tích WoE và giá trị IV, sẽ lựa chọn các biến có ảnh hưởng đáng kể đến mục tiêu dự đoán.
- Thêm bước xác định ngưỡng IV để loại bỏ các biến có giá trị IV dưới ngưỡng quyết định trước (IV < 0.02), là những biến không đủ hữu ích cho dự đoán.

Sau khi lựa chọn được các biến mạnh tiến hành Phân tích tương quan

- Tiến hành phân tích tương quan giữa các biến đã được chọn, nếu phát hiện có sự tương quan cao giữa một số cặp biến, có thể xem xét loại bỏ một trong số chúng (Để lại biến có IV cao nhất trong các cặp).

➢ Sau quá trình áp dụng phân tích WoE – IV và phân tích tương quan đã lựa chọn các biến bao gồm **2 biến Mạnh (Strong)** và **16 biến Yếu (weak)** để xây dựng mô hình.

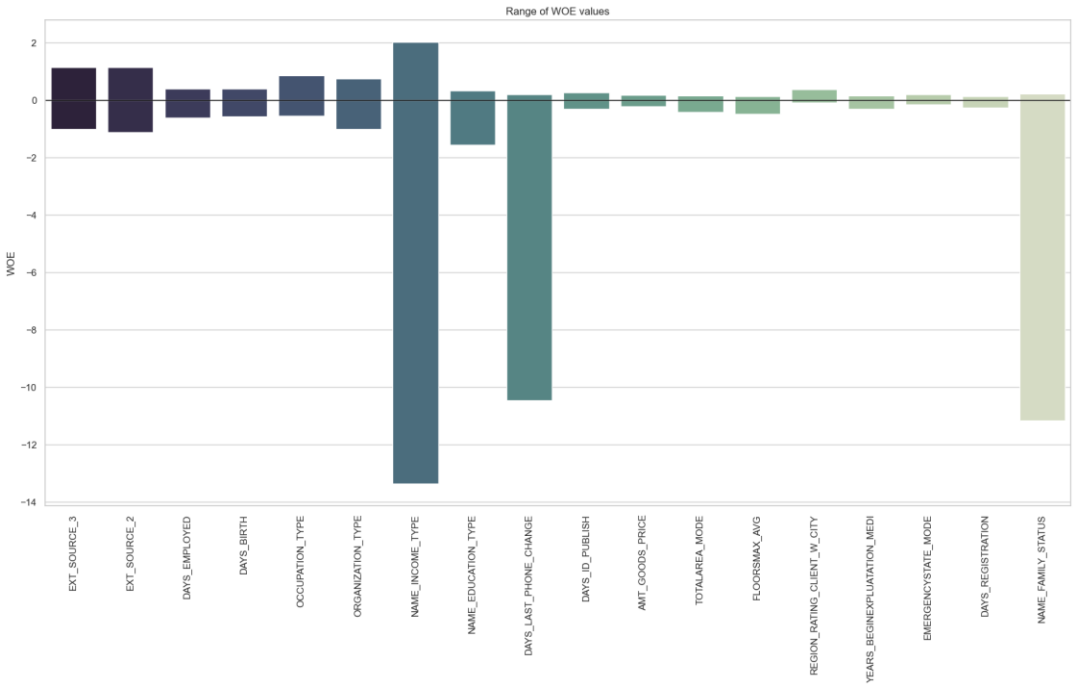
Biểu đồ thể hiện giá trị IV từ cao tới thấp.



Variable Analysis

Sau khi lựa chọn được các biến có IV cao tiến hành phân tích các biến. Kết quả binning được xuất ra file Excel **(Report Variable.xlsx).**

Biểu đồ thể hiện giá trị phạm vi WoE của các biến.



Variable Transformation (WoE)

Mô hình sẽ không hồi quy trực tiếp trên các biến gốc mà thay vào đó giá trị WOE ở từng biến sẽ được sử dụng thay thế để làm đầu vào.

3. Huấn luyện Mô Hình

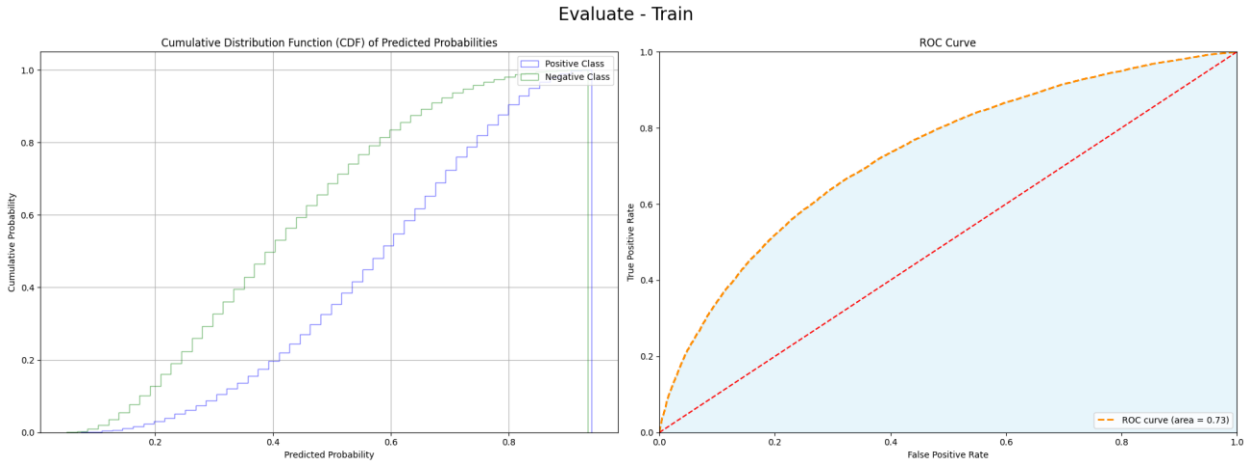
Tiến hành xử lý mất cân bằng ở biến mục tiêu (biến mục tiêu bị mất cân bằng khi biến 1 có nhiều quan sát hơn biến 0), lấy ngẫy nhiên các quan sát 0 bằng với số lượng của quan sát 1.

Chia dữ liệu làm 2 phần với 80% để quá trình huấn luyện và 20% kiểm đỉnh.

Kiểm định mô hình, mô hình sẽ được kiểm định với phương pháp Kolmogorov-Smirnov và chỉ số ROC trên cả 2 tập huấn luyện và kiểm định.

Đối với tập huấn luyện:

- Kolmogorov-Smirnov: 0.3437
- ROC curve: 0.7309



Đối với tập kiểm định:

- Kolmogorov-Smirnov: 0.347
- ROC curve: 0.7354

Evaluate - Test