Dự đoán tình trạng rủi ro nợ xấu vay vốn là vỡ nợ hay không vỡ nợ dựa trên bộ dữ liệu HMEQ (kaggle dataset).

2. Mô hình

Các bước xây dựng mô hình:

- Xóa dữ liệu trùng lặp
- Phân chia tập train / test
- Xử lý các giá trị bị khuyết
- Xử lý các ngoại lệ
- Mã hóa các cột dữ liệu dạng hạng mục
- Xây dựng mô hình

2.1 Xóa dữ liệu trùng lặp (giống bộ dữ liệu credit risk default)

2.2 Phân chia tập train / test

Bộ dữ liệu được tiến hành phân chia trước khi được tiền xử lý tránh tình trạng đánh giá không tự nhiên (công bằng) giữa tập train và test (overfitting).

Sau khi xóa dữ liệu trùng lặp Bộ dữ liệu bao gồm 5960 mẫu dữ liệu (bản ghi) và 13 trường dữ liệu (cột *BAD* là cột nhãn dự đoán).

Phân chia Bộ dữ liệu với tỉ lệ 30% cho dữ liệu test:

- 4768 mẫu dữ liệu train
- 1192 mẫu dữ liệu test

2.3 Xử lý các giá trị bị khuyết

Sau khi phân chia Bộ dữ liệu với tập train và tập test, tiến hành xử lý các giá trị bị khuyết trên tập train và áp dụng biến đổi trên tập test dựa vào các xử lý trên tập train.

- Đối với dữ liệu dạng số: thay thế các giá trị bị khuyết bằng giá trị có khả năng xảy ra nhất, chọn giá trung vị median để thay thế.
- Đối với dữ liệu dạng hạng mục: thay thế bằng hạng mục xuất hiện nhiều nhất.

2.4 Xử lý ngoại lệ - mã hóa dữ liệu hạng mục

- Dữ liệu có xu hướng bị lệch (right-skewed) và có khả năng có nhiễu (outliers), Điều này có thể làm ảnh hưởng chất lượng của mô hình, ngoài ra các trường dữ liệu có khoảng giá trị khác nhau (khoảng lớn, nhỏ) có thể ảnh hưởng đến phương pháp tối ưu cho mô hình (ví dụ hội tụ lâu hơn với gradient descent).

- Sử dụng phương pháp RobustScaler, biến đổi các trường dạng số về cùng một khoảng giá trị, chuẩn hóa dữ liệu về dạng phân phối chuẩn bỏ qua các giá trị nhiễu dựa vào giá trị trung vị median và các khoảng phân vị (tứ phân vị - interquatitle range IQR)

$$value = \frac{value - median}{value\_75th - value\_25th}$$

- Sử dụng mã hóa OneHotEncoder cho các dữ liệu dạng hạng mục.

2.5 Xây dựng mô hình

Sử dụng hai mô hình **Logistic Regression** và **Support Vector Machine** cho tác vụ phân loại nhị phân.
Thử nghiệm với hai trường hợp đầu vào:

- Đầu vào là các đặc trưng đã được tiền xử lý qua các bước ở trên từ bộ dữ liệu gốc.
- Đầu vào là các đặc trưng được mã hóa bằng phương pháp weight of evidence enconder. Sử dụng các giá trị woe trên mỗi bin làm đặc trưng.

2.6 Kết quả

Đặc trưng được tiền xử lý giá trị khuyết, nhiễu và sử dụng OneHotEncoder	
Model	**accuracy (%)**
Logistic Regression	81.62
SVM	86.32

Đặc trưng được tiền xử lý giá trị khuyết và sử dụng WOEEncoder	
Model	**accuracy (%)**
Logistic Regression	85.48
SVM	86.57

Kết quả sử dụng đầu vào là các đặc trưng được xử lý giá trị bị khuyết, ngoại lệ, sử dụng mã hóa OneHotEncoder cho kết quả nhỉnh hơn so với sử dụng đầu vào là các đặc trưng được xử lý giá trị bị khuyết và sử dụng WOEEncoder.