ĐƠN VỊ TỔ CHỨC

ĐỐI TÁC CHIẾN LƯỢC

CUỘC THI

DATA SCIENCE TALENT COMPETITION 2024

VÒNG CHUNG KẾT

Đội thi: Vô Gia Cư

Ngày thực hiện: 25/10/2024

Giới thiệu thành viên

Nguyễn Đức Dũng (leader) - 106042

Nguyễn Văn Hoàng - 106127

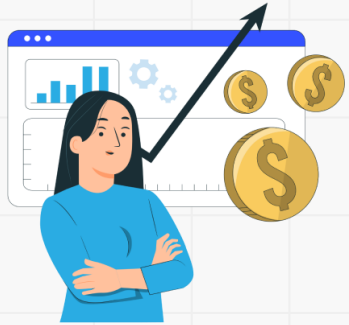
Nguyễn Bảo Phong - 106001

Mục lục

1. Tổng quan

Tổng quan

- Trong lĩnh vực tài chính và ngân hàng, **dữ liệu lịch sử tín dụng** đóng vai trò quan trọng trong việc đánh giá khả năng tín dụng của khách hàng.



Bộ dữ liệu

- **ID cá nhân của khách hàng.**
- **Các biến lịch sử tín dụng.**
- **Label:** Label: 1 - khách hàng bị quá hạn (default), 0 - khách hàng trả nợ đúng hạn.

Nền tảng mô hình

- Ngôn ngữ sử dụng: Python.
- Nền tảng sử dụng: Google Colab.

1. Tổng quan

	Tên cột/ nhóm cột	Mô tả
2	customer_id	Mã định danh của khách hàng
3	label	0: Khách hàng trả nợ đúng hạn 1: Khách hàng quá hạn trả nợ
4	_COUNT_	Số khoản vay theo từng loại ngắn hạn - trung hạn - dài hạn từ ngân hàng - tổ chức phi ngân hàng
5	NUMBER_OF_LOANS_	Tổng tất cả khoản vay (không phân biệt loại) từ ngân hàng - tổ chức phi ngân hàng
6	NUMBER_OF_CREDIT_CARDS_	Tổng số thẻ tín dụng được cấp bởi ngân hàng - tổ chức phi ngân hàng
7	NUMBER_OF_RELATIONSHIP_	Tổng số mối quan hệ tài chính mà khách hàng có với ngân hàng - tổ chức phi ngân hàng
8	NUM_NEW_LOAN_TAKEN_xM	Tổng số khoản vay mới từ ngân hàng - tổ chức phi ngân hàng mà khách hàng thực hiện trong 3 - 6 - 9 - 12 tháng trước đó
9	OUTSTANDING_BAL_LOAN_CURRENT	Số dư nợ của các khoản vay tính đến hiện tại
10	OUTSTANDING_BAL_LOAN_xM	Số dư nợ của các khoản vay trong 3 - 6 - 9 - 12 tháng trước đó
11	OUTSTANDING_BAL_CC_xM	Số dư nợ của thẻ tín dụng trong 3 - 6 - 9 - 12 tháng trước đó
12	OUTSTANDING_BAL_ALL_xM	Tổng số dư nợ cho tất cả các sản phẩm tài chính trong 3 - 6 - 9 - 12 tháng trước đó
13	OUTSTANDING_BAL_LOAN_xM_yM	Số chênh lệch giữa các số dư nợ của các khoản vay trong hai khoảng thời gian
14	OUTSTANDING_BAL_CC_xM_yM	Số chênh lệch giữa các số dư nợ của thẻ tín dụng trong hai khoảng thời gian
15	OUTSTANDING_BAL_ALL_xM_yM	Số chênh lệch giữa tổng các số dư nợ cho tất cả các sản phẩm tài chính trong hai khoảng thời gian
16	INCREASING_BAL_xM_	Số tăng lên trong số dư nợ của các khoản vay - thẻ tín dụng - tất cả sản phẩm tài chính trong 3 - 6 tháng
17	OUTSTANDING_BAL_CC_CURRENT	Số dư nợ của thẻ tín dụng tính đến hiện tại
18	CREDIT_CARD_MONTH_SINCE_xDPD	Số tháng kể từ khi khách hàng có thanh toán quá hạn 10 - 30 - 60 - 90 ngày trên khoản thanh toán thẻ tín dụng gần nhất
19	CREDIT_CARD_NUMBER_OF_LATE_PAYMENT	Tổng số lần thanh toán trễ của khách hàng trên các thẻ tín dụng
20	ENQUIRIES_FROM_FOR_xM	Số lượt tra cứu tín dụng liên quan đến các khoản vay - thẻ tín dụng từ ngân hàng - tổ chức phi ngân hàng trong 3 - 6 - 9 - 12 tháng trước đó
21	ENQUIRIES_FROM_xM_yM	Số lượt tra cứu tín dụng liên quan đến các sản phẩm tài chính từ ngân hàng - tổ chức phi ngân hàng trong 2 khoảng thời gian
22	OUTSTANDING_BAL_ALL_CURRENT	Tổng số dư nợ cho tất cả các sản phẩm tài chính tính đến hiện tại

2. Phân tích dữ liệu

	0
customer_id	0
label	0
SHORT_TERM_COUNT	2000
MID_TERM_COUNT	2000
LONG_TERM_COUNT	2000
...	...
ENQUIRIES_FROM_NON_BANK_6M_9M	2000
ENQUIRIES_FROM_NON_BANK_9M_12M	2000
ENQUIRIES_FROM_NON_BANK_6M_12M	2000
ENQUIRIES_FROM_NON_BANK_3M_12M	2000
OUTSTANDING_BAL_ALL_CURRENT	2000

Dữ liệu trống

SHORT_TERM_COUNT	SHORT_TERM_COUNT_BANK	SHORT_TERM_COUNT_NON_BANK
1.0	1.0	1.0
1 + 10.0	= 7.0	+ 4.0
7.0	7.0	1.0

Lỗi Logic

CREDIT_CARD_MONTH_SINCE_60DPD	CREDIT_CARD_MONTH_SINCE_90DPD
431.0	431.0
431.0	431.0
431.0	431.0
431.0	431.0
431.0	431.0
...	...
431.0	431.0
431.0	431.0
431.0	431.0
431.0	431.0
431.0	431.0

Lỗi trùng lặp dữ liệu

SAU QUÁ TRÌNH XỬ LÝ SƠ BỘ

SHORT_TERM_COUNT	8
MID_TERM_COUNT	7
LONG_TERM_COUNT	7
SHORT_TERM_COUNT_BANK	8

Lượng các ô trống ở các cột count chỉ còn chiếm 0.5% so với 20000 dữ liệu ban đầu.

2. Phân tích dữ liệu

Dữ liệu giá trị số đếm

NUMBER_OF_LOANS_BANK
1.0
7.0
10.0
1.0
1.0
...
4.0
1.0
4.0
10.0
1.0

Dữ liệu giá trị số tiền

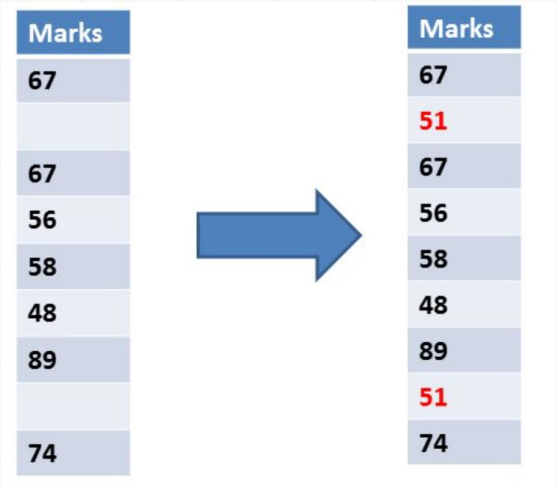
OUTSTANDING_BAL_CC_9M_12M
1.000000e+06
1.000013e+06
1.000013e+06
1.000013e+06
1.000013e+06
...
9.999900e+05
1.000000e+06
1.000010e+06
1.000011e+06
1.000011e+06

Cùng kiểu dữ liệu -> cộng

Khác kiểu dữ liệu -> nhân

Tạo ra tổng cộng **713 đặc trưng**
(bao gồm các đặc trưng ban đầu)

2. Phân tích dữ liệu



713 đặc trưng

Fill theo trung vị | **Standard Scaler**

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

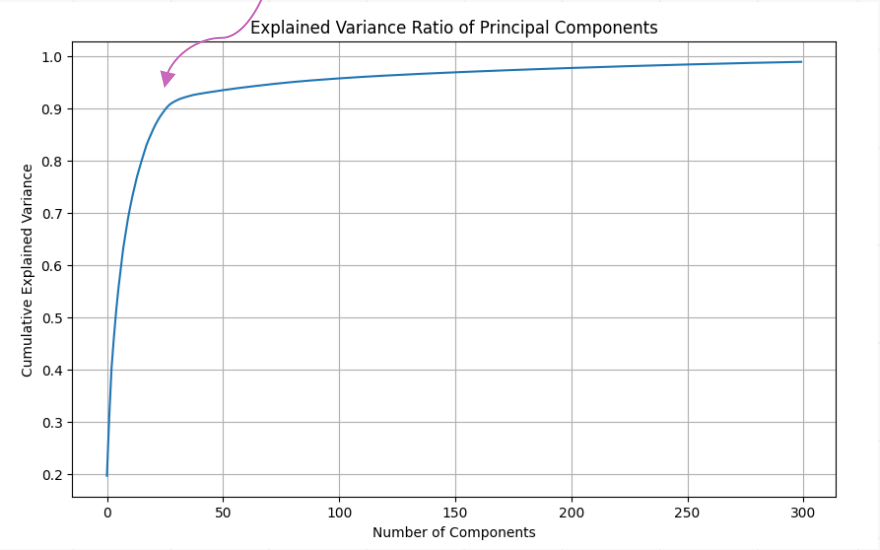
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Kỹ thuật PCA

Giảm xuống còn **30 thành phần chính** để đánh giá

2. Phân tích dữ liệu

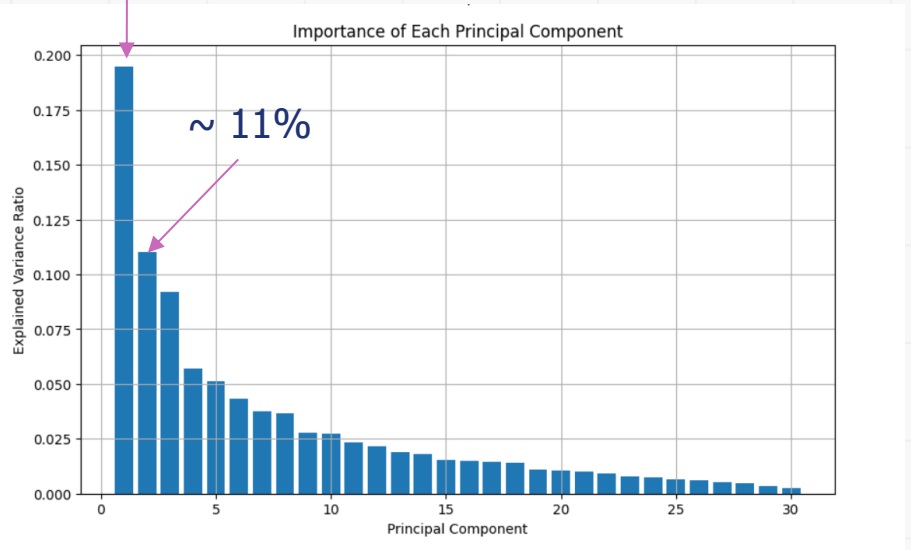
elbow point



Việc chọn 30 PCs vẫn giữ được trên **90%** dữ liệu

~ 19%

~ 11%



Chọn các đặc trưng đầu được xem xét do chứa **lượng lớn** dữ liệu gốc.

2. Phân tích dữ liệu

Tìm các đặc trưng có **ảnh hưởng lớn** với từng PC, nhằm xác định một số **tương quan có thể hữu ích.**

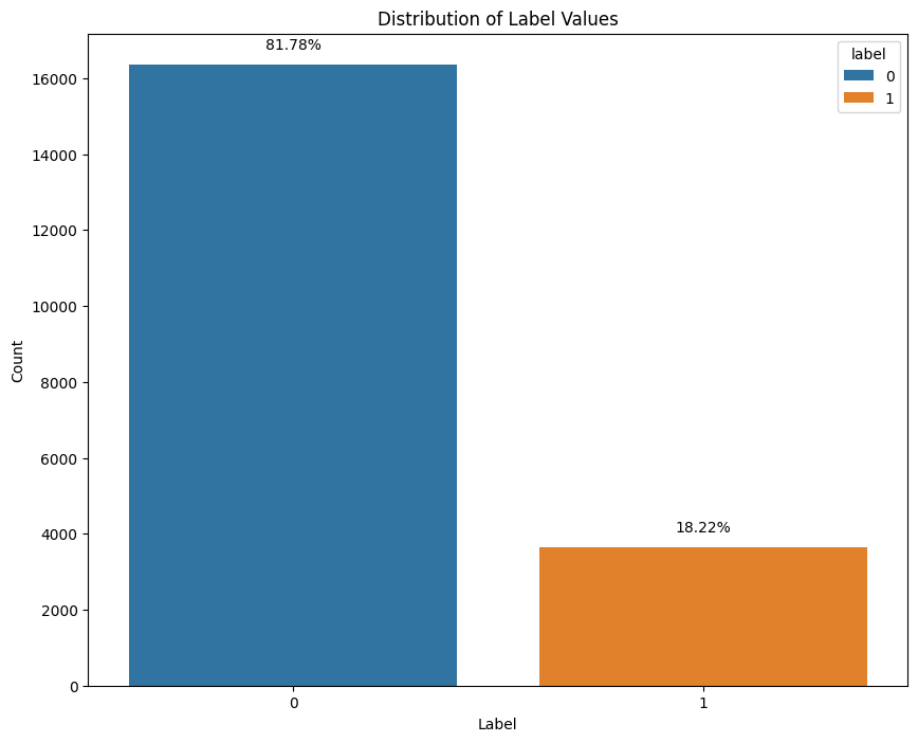
```
Top 5 features for PC1:
  - NUMBER_OF_LOANS: 0.0744
  - LONG_TERM_COUNT_NON_BANK_x_NUMBER_OF_LOANS: 0.0744
  - SHORT_TERM_COUNT_x_MID_TERM_COUNT: 0.0742
  - LONG_TERM_COUNT_x_NUMBER_OF_LOANS: 0.0740
  - SHORT_TERM_COUNT_x_NUMBER_OF_LOANS: 0.0733
--------------------
Top 5 features for PC2:
  - ENQUIRIES_9M_12M_x_ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0727
  - ENQUIRIES_3M_12M: 0.0726
  - ENQUIRIES_6M_9M_x_ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0725
  - ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0716
  - LONG_TERM_COUNT_NON_BANK_x_ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0714
--------------------
Top 5 features for PC3:
  - ENQUIRIES_FROM_NON_BANK_3M_x_ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0826
  - ENQUIRIES_FROM_NON_BANK_FOR_CC_3M_x_ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0815
  - ENQUIRIES_FROM_NON_BANK_FOR_CC_6M_x_ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0789
  - ENQUIRIES_FROM_NON_BANK_3M_x_ENQUIRIES_FROM_NON_BANK_6M_9M: 0.0785
  - ENQUIRIES_FROM_NON_BANK_3M_x_ENQUIRIES_6M_9M: 0.0784
--------------------
```

```
Top 5 features for PC2:
  - ENQUIRIES_9M_12M_x_ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0727
  - ENQUIRIES_3M_12M: 0.0726
  - ENQUIRIES_6M_9M_x_ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0725
  - ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0716
  - LONG_TERM_COUNT_NON_BANK_x_ENQUIRIES_FROM_NON_BANK_3M_12M: 0.0714
--------------------
```

Xem xét thử mối quan hệ giữa các ENQUIRIES_xM_yM.

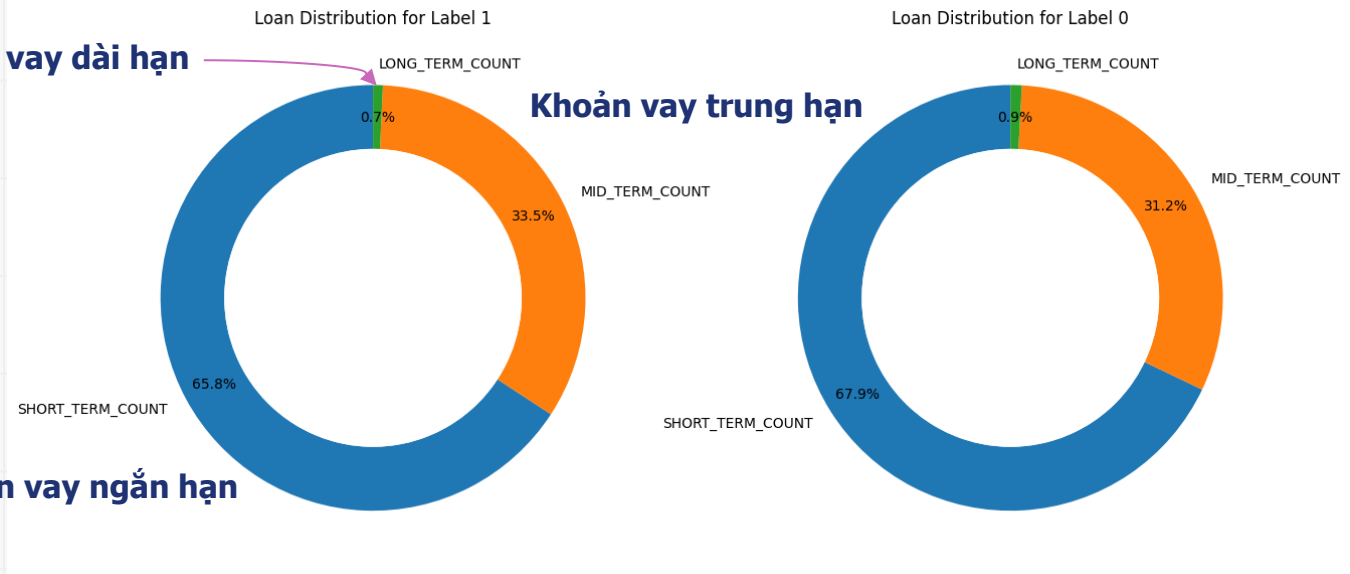
3. Khai phá dữ liệu

Phân bố của label



3. Khai phá dữ liệu

Khoản vay ngắn, trung và dài hạn



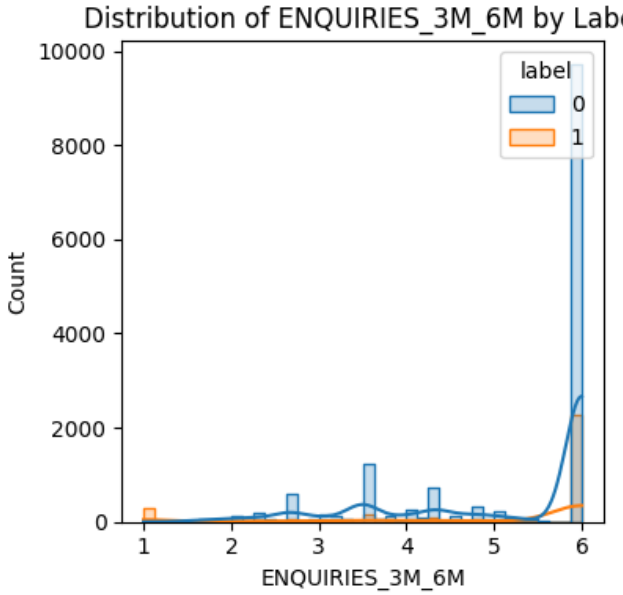
Khoản vay dài hạn

Khoản vay trung hạn

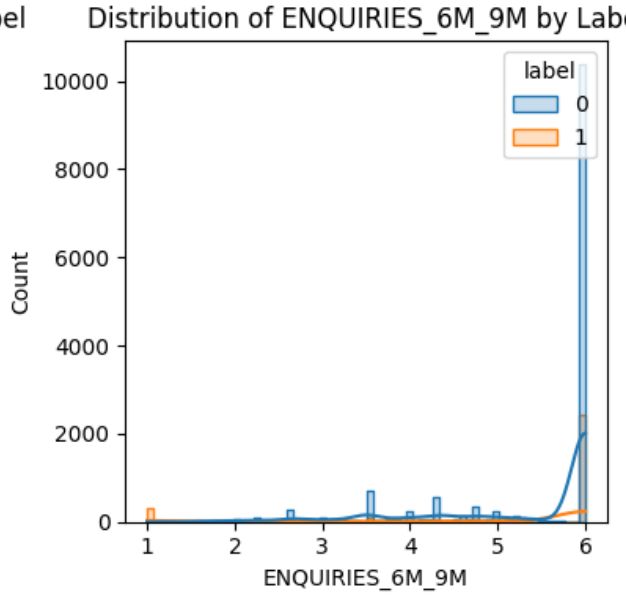
Khoản vay ngắn hạn

3. Khai phá dữ liệu

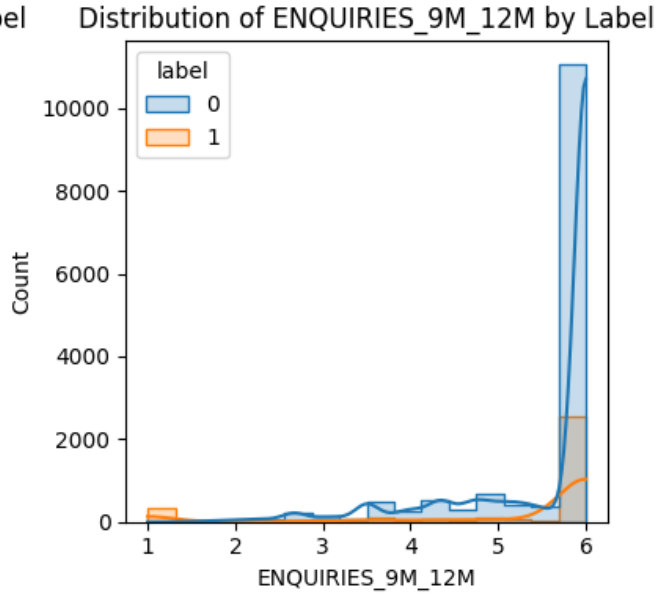
Số lượt tra cứu tín dụng liên quan đến các sản phẩm tài chính



Số lượt tra cứu 3-6 tháng trước



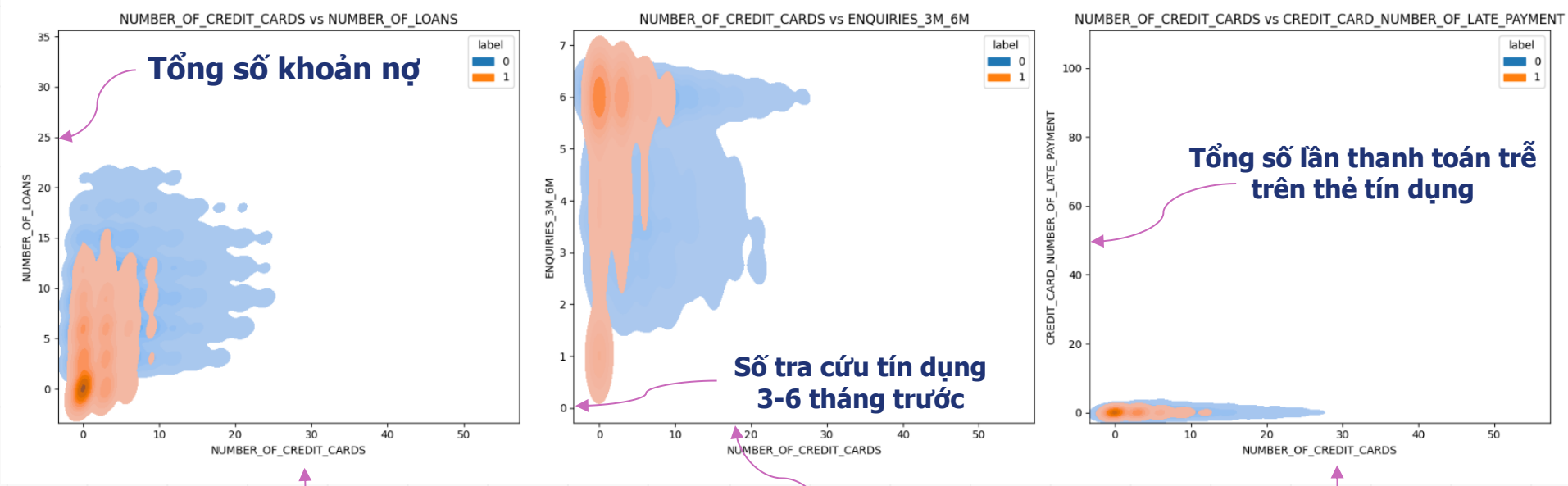
Số lượt tra cứu 6-9 tháng trước



Số lượt tra cứu 9-12 tháng trước

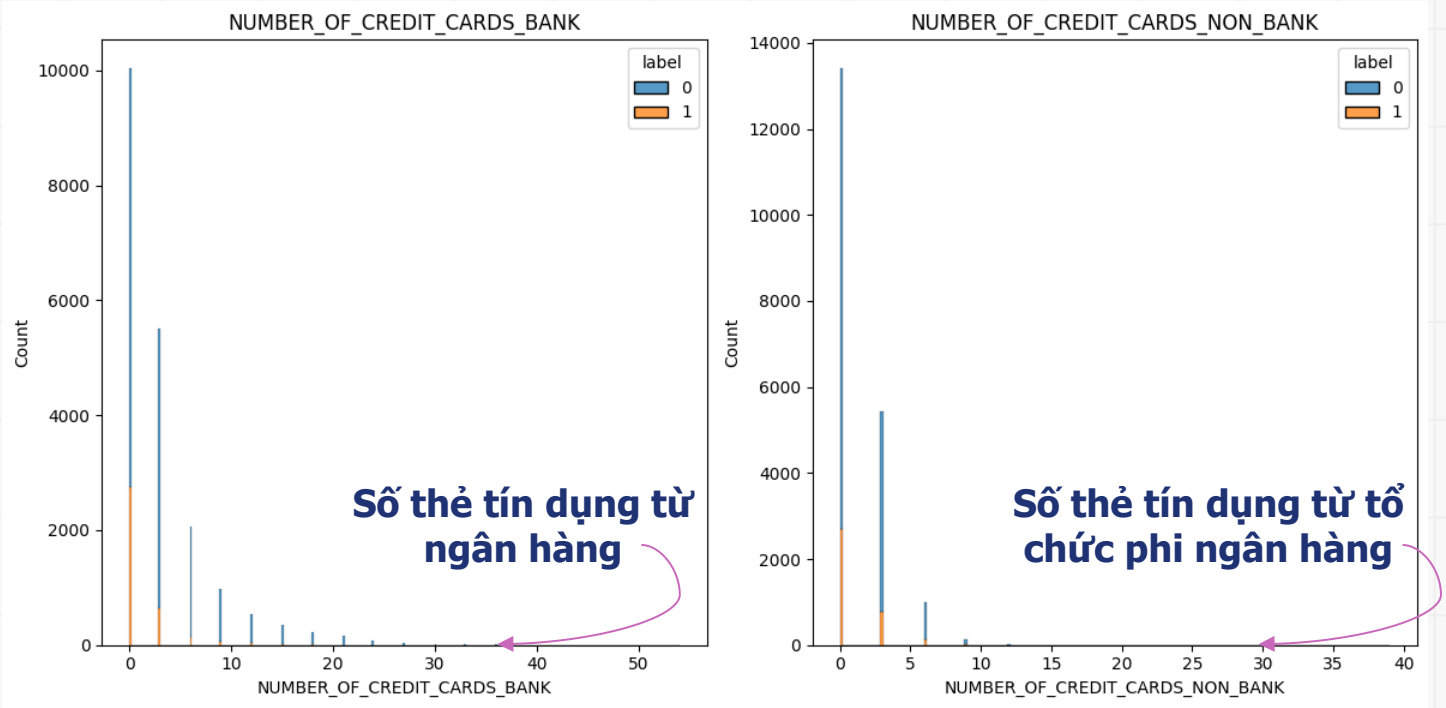
3. Khai phá dữ liệu

Tổng số thẻ tín dụng



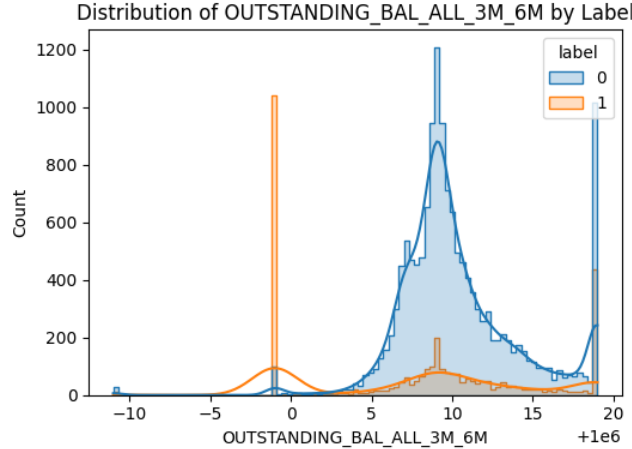
3. Khai phá dữ liệu

Số thẻ tín dụng được cấp bởi ngân hàng và tổ chức phi ngân hàng

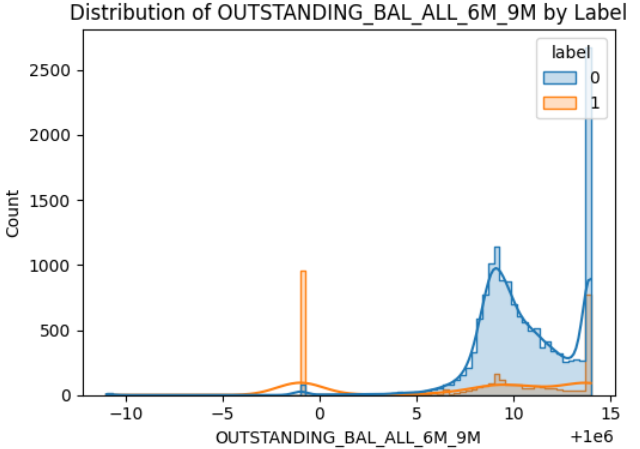


3. Khai phá dữ liệu

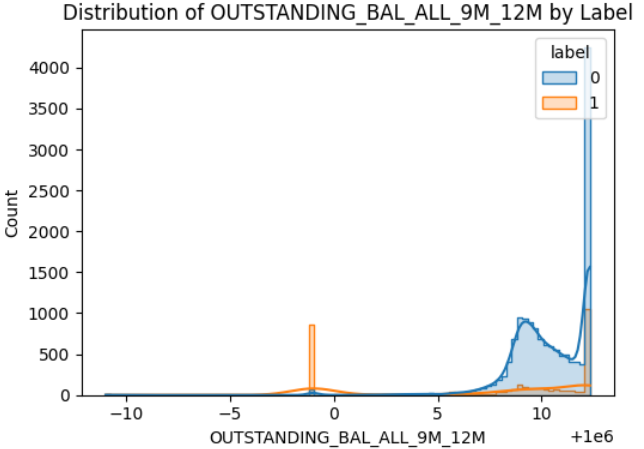
Chênh lệch số dư nợ theo thời gian



Chênh lệch dư nợ 3-6 tháng



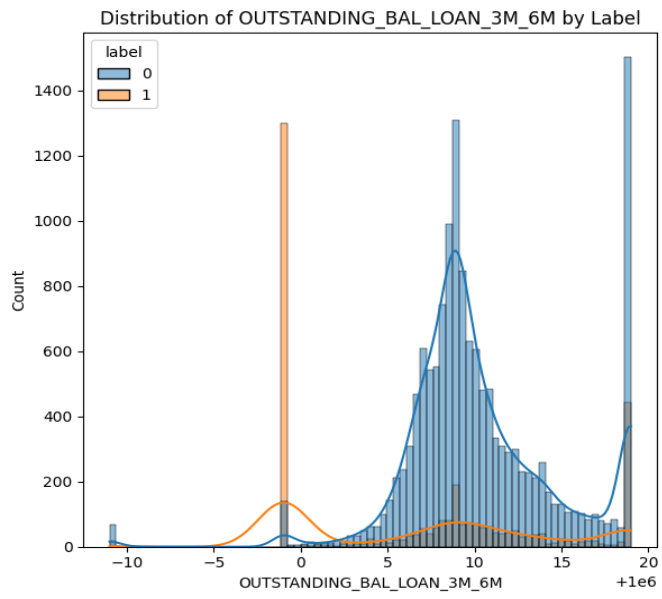
Chênh lệch dư nợ 6-9 tháng



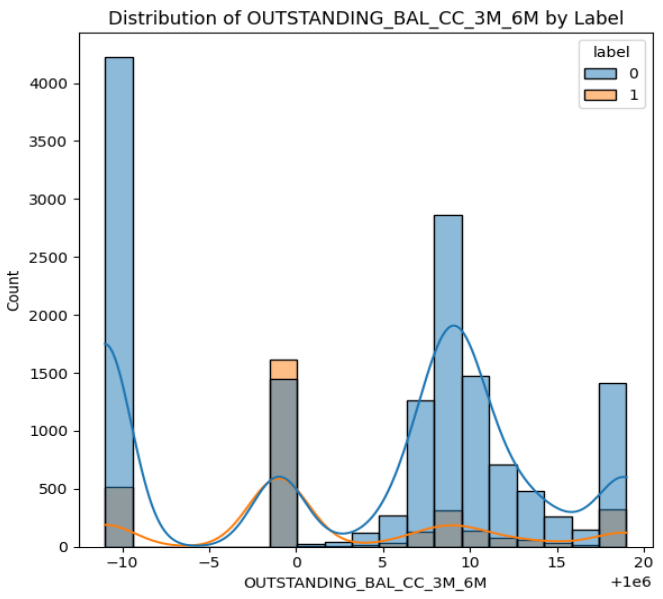
Chênh lệch dư nợ 9-12 tháng

3. Khai phá dữ liệu

Chênh lệch số dư nợ của khoản vay và thẻ tín dụng



Chênh lệch dư nợ khoan vay 3-6 tháng



Chênh lệch dư nợ thẻ tín dụng 3-6 tháng

3. Khai phá dữ liệu

Không trả nợ đúng hạn

- Ít lượt tra cứu tín dụng (1-2 lượt)
- Ít thẻ tín dụng (<10)
- Số dư nợ tăng lên theo thời gian

Trả nợ đúng hạn

- Nhiều lượt tra cứu tín dụng
- Số lượng thẻ tín dụng nhiều
- Số dư nợ giảm dần theo thời gian

4. Mô hình dự đoán

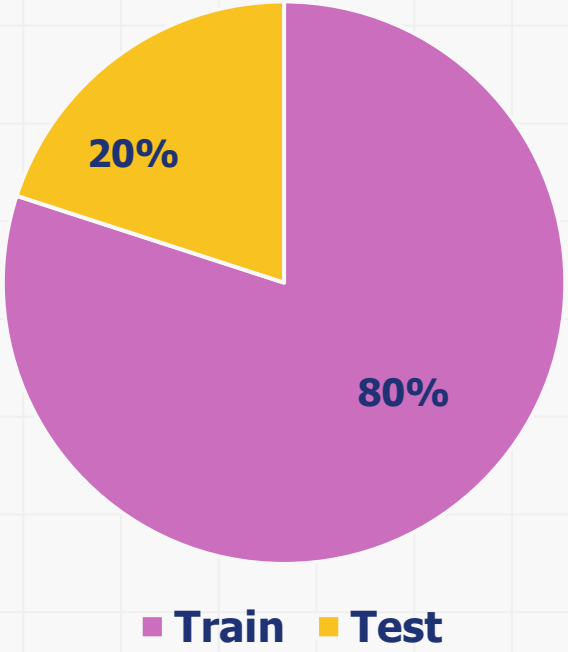
Xây dựng giải pháp

Lựa chọn mô hình

Đánh giá mô hình

4.1. Xây dựng giải pháp

DATASET



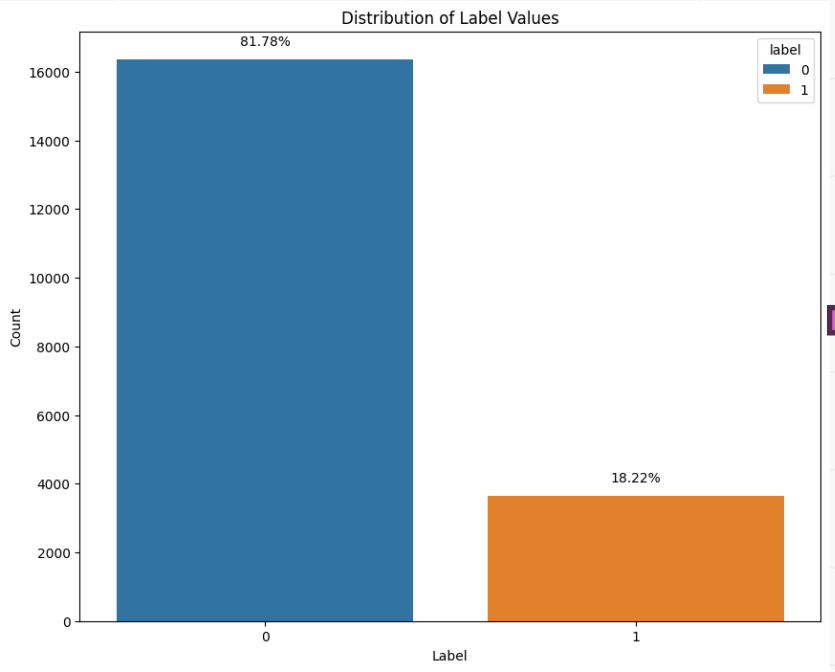
STANDARD SCALER

$$z = \frac{x - \mu}{\sigma}$$

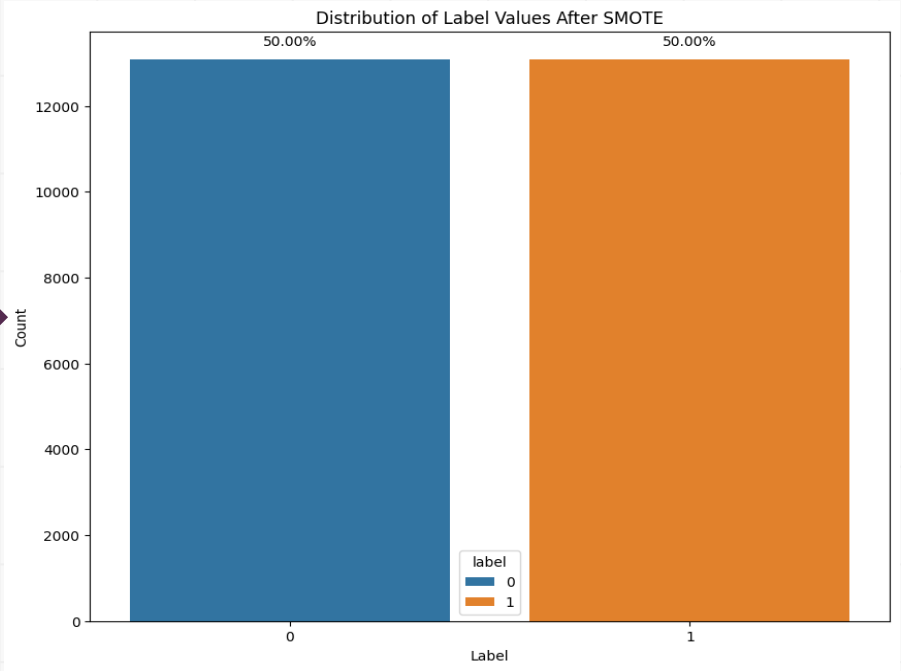
$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

4.1. Xây dựng giải pháp



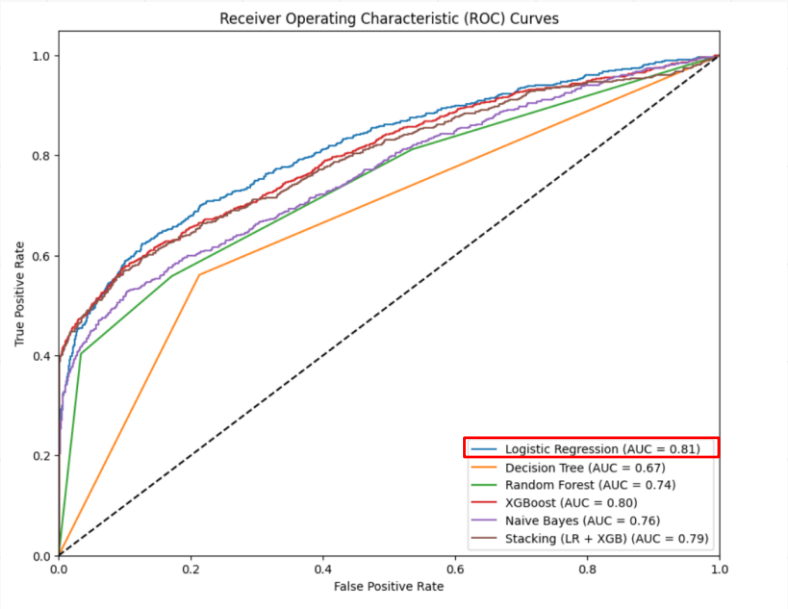
SMOTE



4.2. Đánh giá giải pháp

Đánh giá các chỉ số trên mô hình

	Model	AUC	Gini	KS
0	Logistic Regression	0.814702	0.629404	0.496510
1	Decision Tree	0.674132	0.348264	0.348264
2	Random Forest	0.743630	0.487261	0.387403
3	XGBoost	0.800865	0.601730	0.477691
4	Naive Bayes	0.762342	0.524684	0.425400
5	Stacking (LR + XGB)	0.792486	0.584972	0.471138



⇨ **Lựa chọn mô hình Logistic Regression**

4.2. Đánh giá giải pháp

ƯU ĐIỂM

- Phân biệt 2 nhóm khách hàng rất tốt.
- Dễ hiểu, dễ triển khai trong doanh nghiệp, và tính ổn định cao.

NHƯỢC ĐIỂM

- Kém hiệu quả với các đặc trưng đa cộng tuyến.

-> Sử dụng PCA giúp tránh đa cộng tuyến giữa các đặc trưng

5. Ứng dụng thực tiễn

SCORECARD

$$Score = Offset + Factor * \ln(p/(1-p))$$

Base Point – Factor * ln(p’/(1-p’))

Chênh lệch / ln(2)

5. Ứng dụng thực tiễn

SCORECARD

Phong

- Truy cập vào tín dụng trong 3 tháng gần đây 3 lần, 12 tháng là 17 lần.
- Tổng số dư nợ trên mọi nền tảng là 170tr.
- Có 1 thẻ tín dụng.

362.5

386

Dũng

- Truy cập vào tín dụng trong 3 tháng là 7 lần, 12 tháng gần đây 72 lần.
- Tổng số dư nợ trên mọi nền tảng là 100tr.
- Có 7 thẻ tín dụng.

411.36

5. Ứng dụng thực tiễn

Thu thập và khám phá dữ liệu

- Thu thập dữ liệu khách hang
- Tiền xử lý
- Trực quan hoá

Huấn luyện mô hình

- Chọn mô hình phù hợp
- Đánh giá hiệu suất

Tính ScoreCard cho từng khách hang

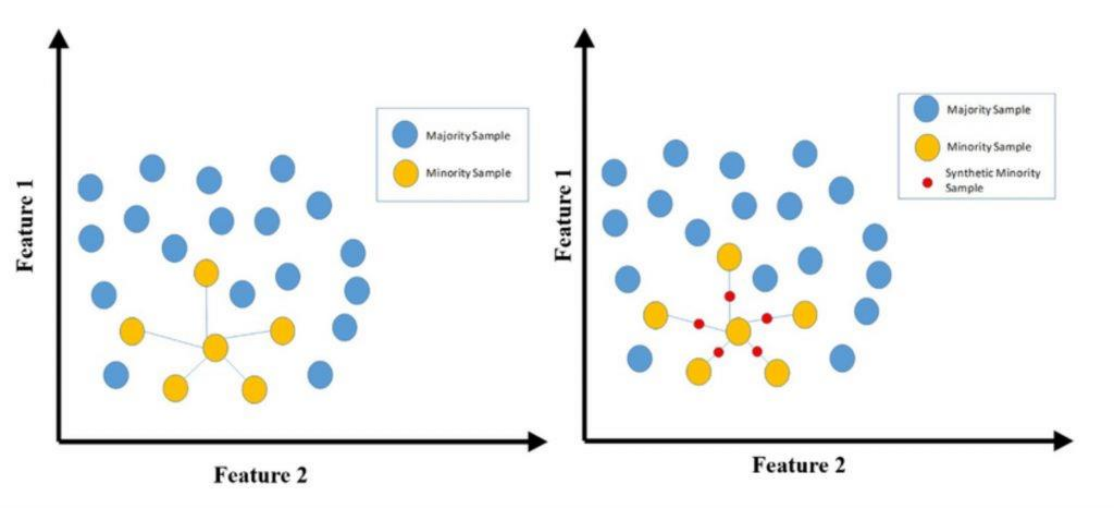
- Log-Odds
- Sử dụng Factor và Offset để tinh chỉnh lại

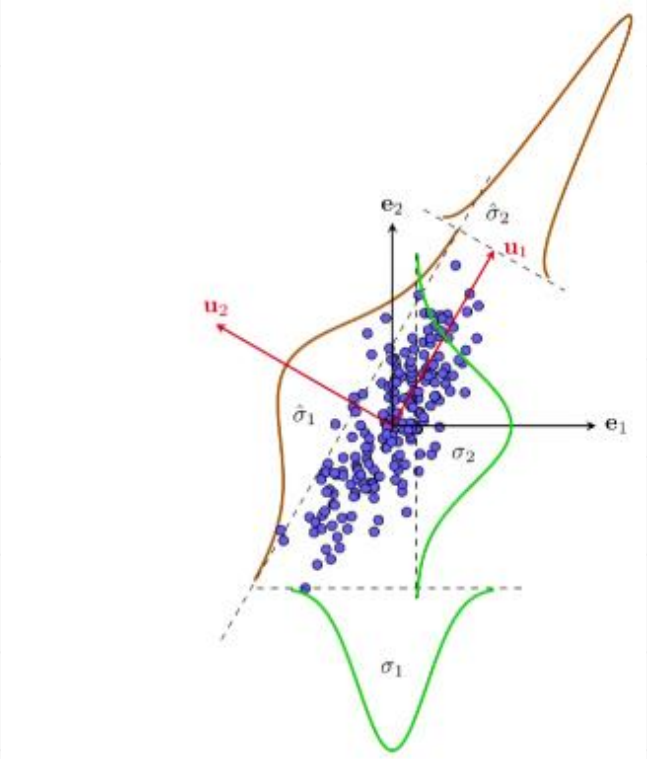
Tích hợp vào trang web quản lý khách hàng

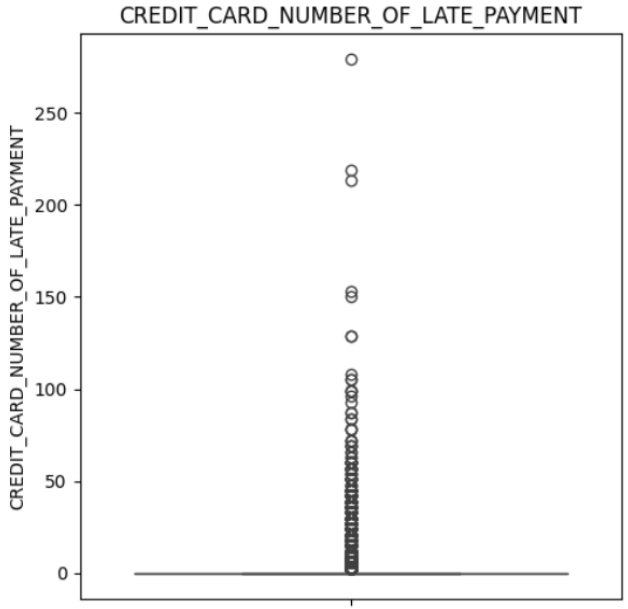
Điều chỉnh chính sách cho vay, phân cụm, cá nhân hoá dịch vụ tài chính

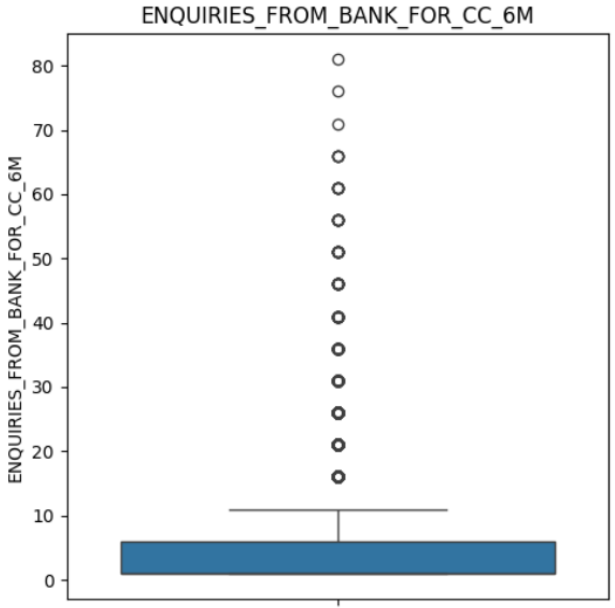
- **Luôn cập nhật dữ liệu thường xuyên và đánh giá hiệu suất định kì**
- **Thay đổi biến nếu có thể cải thiện hiệu suất mô hình**

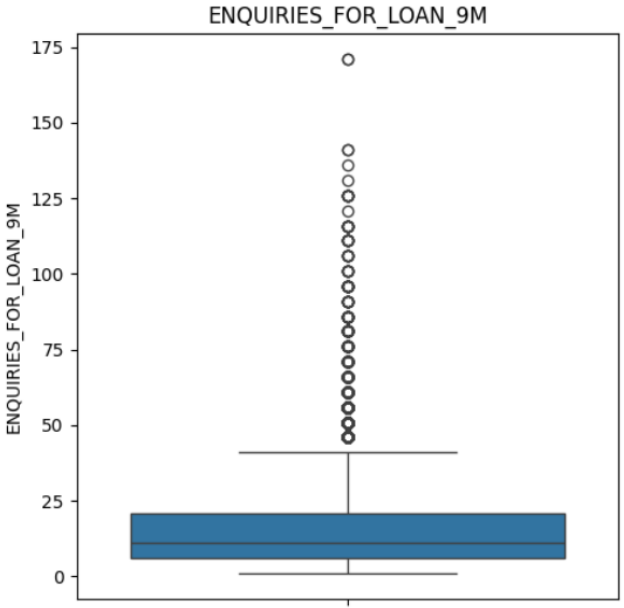

THANK
YOU!


Feature 1
Feature 2
Majority Sample
Minority Sample
Feature 1
Feature 2
Majority Sample
Minority Sample
Synthetic Minority Sample


$\mathbf{e}_2$
$\hat{\sigma}_2$
$\mathbf{u}_1$
$\mathbf{u}_2$
$\hat{\sigma}_1$
$\mathbf{e}_1$
σ_2
σ_1


CREDIT_CARD_NUMBER_OF_LATE_PAYMENT
CREDIT_CARD_NUMBER_OF_LATE_PAYMENT
0
50
100
150
200
250


ENQUIRIES_FROM_BANK_FOR_CC_6M
ENQUIRIES_FROM_BANK_FOR_CC_6M
0
10
20
30
40
50
60
70
80


ENQUIRIES_FOR_LOAN_9M
ENQUIRIES_FOR_LOAN_9M
0
25
50
75
100
125
150
175