BÁO CÁO BÀI THI VÒNG 4
ĐỘI THI VÔ GIA CƯ

1. Tổng quan

1.1 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.

- Đội thi sẽ phân tích và xử lý dữ liệu lịch sử tín dụng của khách hàng. Sau khi tiền xử lý dữ liệu và trực quan hoá dữ liệu đó, chúng em sẽ hiểu hơn về những khách hàng có khả năng trả nợ đúng hạn và không có khả năng trả nợ đúng hạn, từ đó đưa ra mô hình dự đoán cho hai tệp khách hàng đó và đề xuất ứng dụng cho thực tiễn.

1.2 Bộ dữ liệu- Nền tảng mô hình

- ID cá nhân của khách hàng.

- Các biến lịch sử tín dụng : lịch sử vay nợ, khoản vay hiện tại, dư nợ, số lần quá hạn trả nợ và mức độ quá hạn, các khoản vay có đảm bảo và không đảm bảo, số lần tra cứu thông tin tín dụng (chi tiết xem hình sau).

	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

- Label: Label: 1 - khách hàng bị quá hạn (default), Label: 0 - khách hàng trả nợ đúng hạn.
- Ngôn ngữ sử dụng: Python
- Nền tảng sử dụng: Google Colab, PowerBI,…

2. Phân tích sơ bộ dữ liệu

2.1 Làm sạch dữ liệu theo Logic

- Các dữ liệu BTC cung cấp là nguồn dữ liệu thô, cần qua quá trình xử lý trước khi đưa vào phân tích.

- Ngoại trừ dữ liệu về cá nhân khách hàng, label; các dữ liệu còn lại thuộc 1 trong 2 kiểu chính:
  + Dữ liệu giá trị số đếm, ví dụ: Số lượng các khoản vay, thẻ tín dụng, … theo từng loại, số lần tra cứu, …
  + Dữ liệu giá trị số tiền, ví dụ: Số dư nợ theo từng loại, số chênh lệch giữa các tháng, …

- Một số vấn đề trong nguồn dữ liệu (có thể do yếu tố khách quan hoặc chủ quan):
  + Giá trị ô dữ liệu bị trống ('NaN'): Mỗi cột đều có tới 10% ô trống (2000/20000 ô).

	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

+ Lỗi sai logic (ví dụ: _COUNT_BANK + _COUNT_NON_BANK = _COUNT + 1).

SHORT_TERM_COUNT	SHORT_TERM_COUNT_BANK	SHORT_TERM_COUNT_NON_BANK
1.0	1.0	1.0
10.0	7.0	4.0
7.0	7.0	1.0

+ Một số cột dữ liệu xuất hiện lượng lớn (gần như hoàn toàn) cùng một giá trị. (ví dụ các cột liên quan đến CREDIT_CARD_MONTH…)

CREDIT_CARD_MONTH_SINCE_10DPD	CREDIT_CARD_MONTH_SINCE_30DPD	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
NaN	431.0	431.0	431.0
431.0	431.0	431.0	431.0
...	...	...	...
431.0	NaN	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

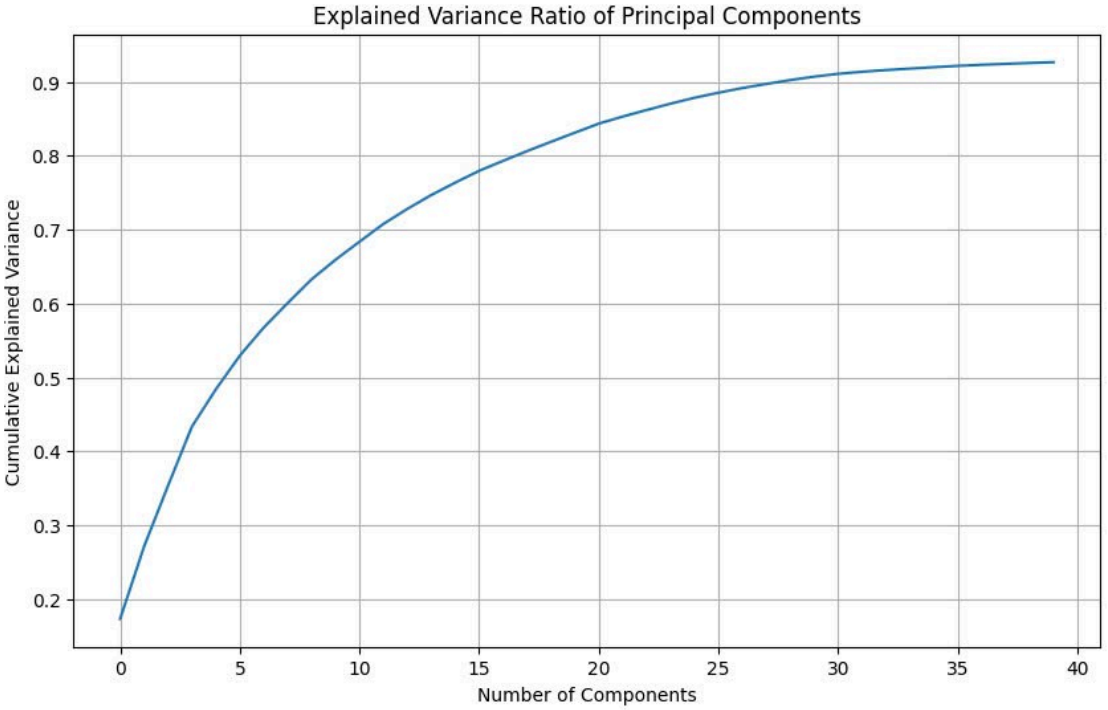
- Từ đó, đội thi đã đưa ra các giải pháp để xử lý sơ bộ dữ liệu như sau (thứ tự thực hiện các quá trình trên sẽ được đề cập sau).
  + Điền giá trị các ô trống theo logic đã nghiên cứu (chủ yếu các giá trị rời rạc), và thực hiện điều chỉnh toàn bộ để logic đúng.
    Ví dụ: Các giá trị của mỗi ô dữ liệu (trừ 2 cột đầu) đều giảm 1 đơn vị. Lý giải là do có thể ban đầu dữ liệu được tăng lên 1 để phân biệt 0 và NaN.
  + Xóa 4 cột liên quan CREDIT_CARD_MONTH_SINCE_ … do các ô đều xuất hiện 431. Theo dự đoán của đội thi, có thể là lỗi HTML 431 Request Header Fields Too Large. Hoặc theo 1 cách nào đó, tuy nhiên dữ liệu trong cột này đều không có ý nghĩa.
  + Sau quá trình trên, lượng các ô trống ở các cột COUNT chỉ chiếm <2%, (không đáng kể)

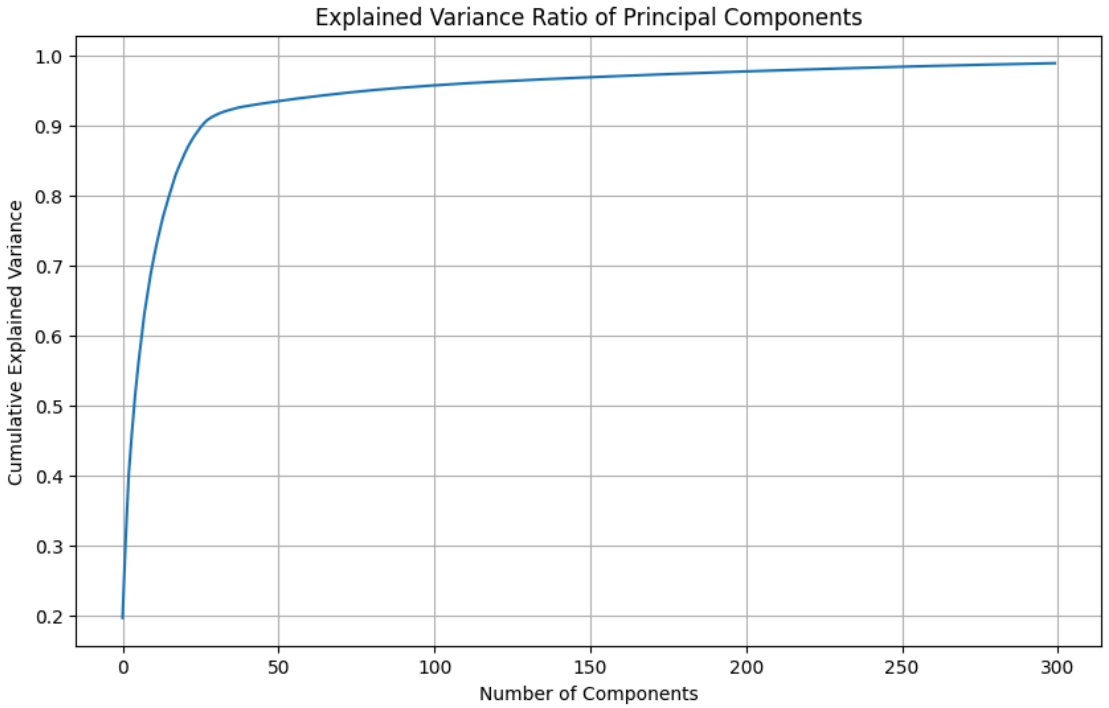
SHORT_TERM_COUNT	8
MID_TERM_COUNT	7
LONG_TERM_COUNT	7
SHORT_TERM_COUNT_BANK	8

+ Lý do không fill các cột khác: Do các cột khác, ta rất khó để đưa ra một số công thức logic giữa chúng, nên ta sẽ không vội vàng fill dữ liệu để tránh mất tính thực tế.
+ Cụ thể, ta chỉ cần fill dữ liệu ở phần Feature Creation trước khi bước vào PCA.

2.2 Phát triển đặc trưng

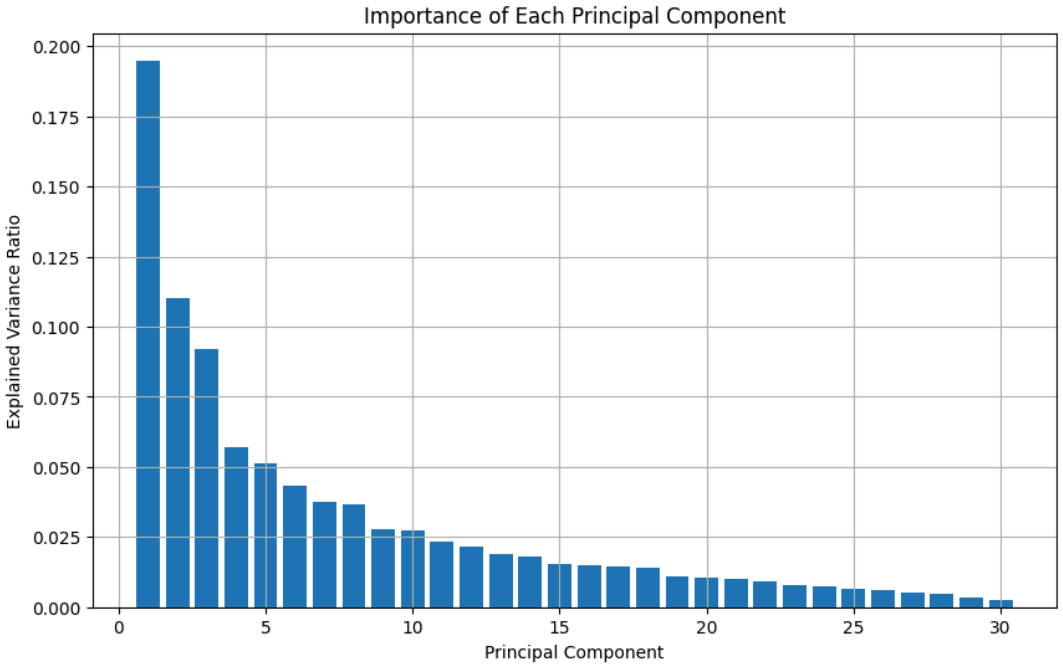
- Chúng ta sẽ bỏ những đặc trưng có tương quan cao với nhau (>0.8), chỉ chọn 1 trong số đó, để đơn giản hoá mô hình và sử dụng 35 đặc trưng còn lại để tạo đặc trưng mới.
- Với những đặc trưng có cùng kiểu dữ liệu, cùng là giá trị rời rạc, giá trị thấp - như số lượng thẻ tín dụng, khoản vay, … hay cùng là giá trị liên tục, giá trị cao - như tổng số dư nơ, chênh lệch dư nợ giữa các tháng… sẽ được cộng với nhau để tạo ra đặc trưng mới. Nếu các giá trị khác biệt, ta thực hiện nhân với nhau để đảm bảo sự tương quan giữa các đặc trưng gốc. Ta tạo ra được 713 đặc trưng theo cách này.
- Ta thực hiện fill dữ liệu trống bằng trung vị của các giá trị trong đặc trưng đó, và kết hợp sử dụng Standard Scaler đưa bộ dữ liệu về [0,1] nhằm đơn giản hóa cho các kỹ thuật mà không mất tính tổng quan.
- Sau đó, ta thực hiện khảo sát để lựa chọn n hợp lý: (n là số chiều giảm xuống sau khi dùng PCA)





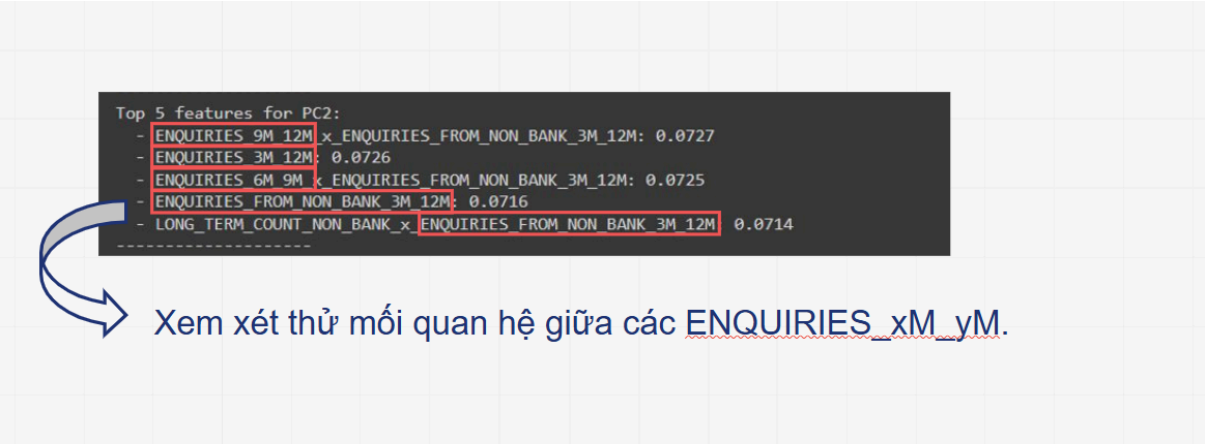
- Ta chọn elbow point trong đồ thị trên, khoảng tầm 25-35, đặc biệt khi kiểm tra với trường hợp nhỏ hơn, ta có thể thấy rõ n = 30 là hợp lý để vừa đủ giữ > 90% dữ liệu ban đầu.

- Ta đi tìm hiểu các features chính ảnh hưởng đến các PC đã liệt kệ (khoảng 5 features), rồi dựa trên đó để đưa ra một số tương quan nhất định, tiếp tục bước vào phần EDA (Khai phá dữ liệu).



- Ta sẽ vẽ ra bảng để xem xét từng PCs mang theo bao nhiêu phần trăm dữ liệu. Dựa vào bảng trên, ta sẽ chọn 5 PCs đầu tiên để nghiên cứu và xem xét

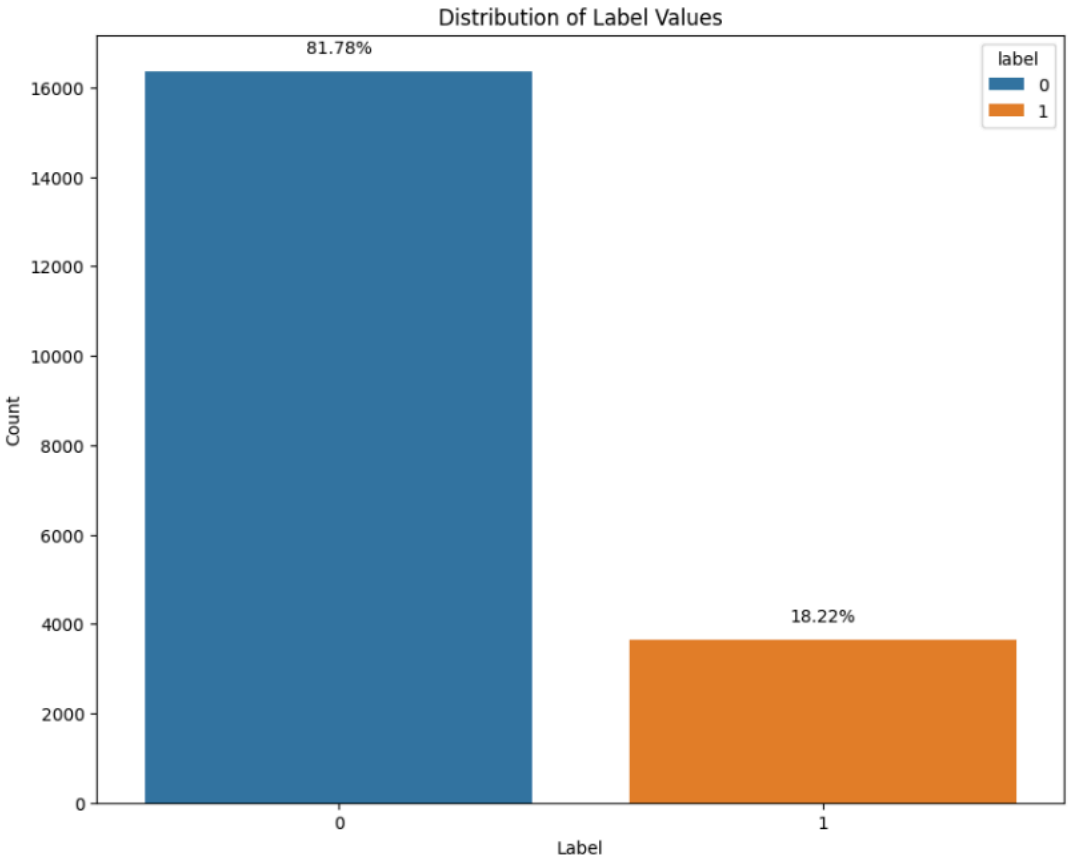
```
Top 5 features for PC1:
  - NUMBER_OF_LOANS: 0.0744
  - LONG_TERM_COUNT_NON_BANK_x_NUMBER_OF_LOANS: 0.0744
  - SHORT_TERM_COUNT_x_MID_TERM_COUNT: 0.0742
  - LONG_TERM_COUNT_x_NUMBER_OF_LOANS: 0.0740
  - SHORT_TERM_COUNT_x_NUMBER_OF_LOANS: 0.0733
--------------------
Top 5 features for PC2:
  - ENQUIRIES_9M_12M_x_ENQUIRIES_FROM_NON_BANK_3M_12M: 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
--------------------
```



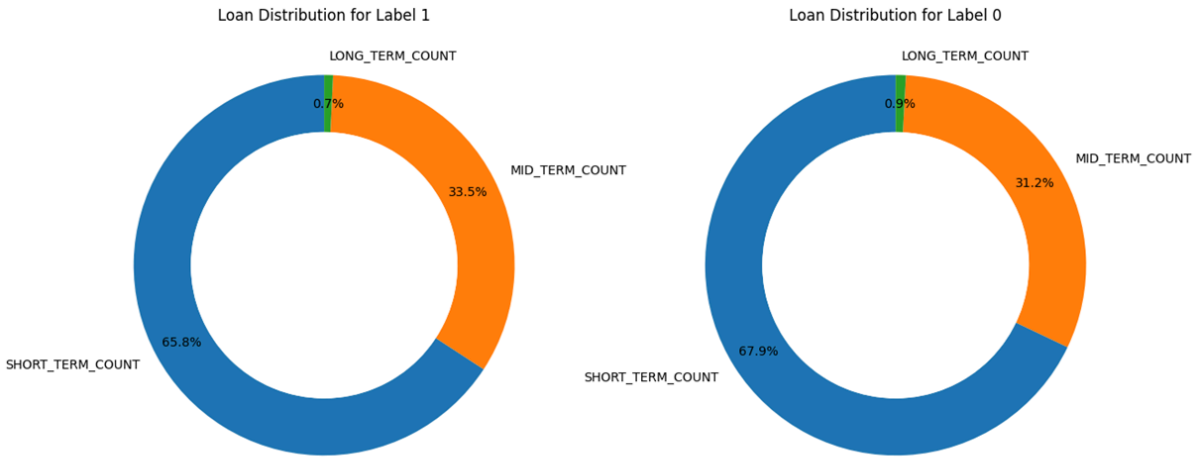
+ Ở đây ta còn phát hiện thêm rằng: tầm quan trọng của đặc trưng ENQUIRIES_9M_12M khi kết hợp với ENQUIRIES_ FROM_ NON_ BANK_ 3M_12M đã tăng lên so với bản thân đặc trưng gốc.

3. Khai phá dữ liệu

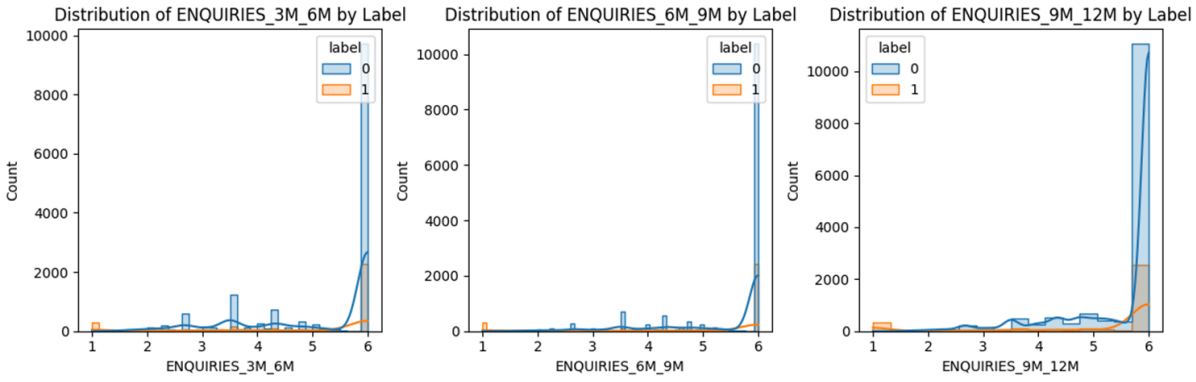
3.1 Với bộ dữ liệu ban đầu



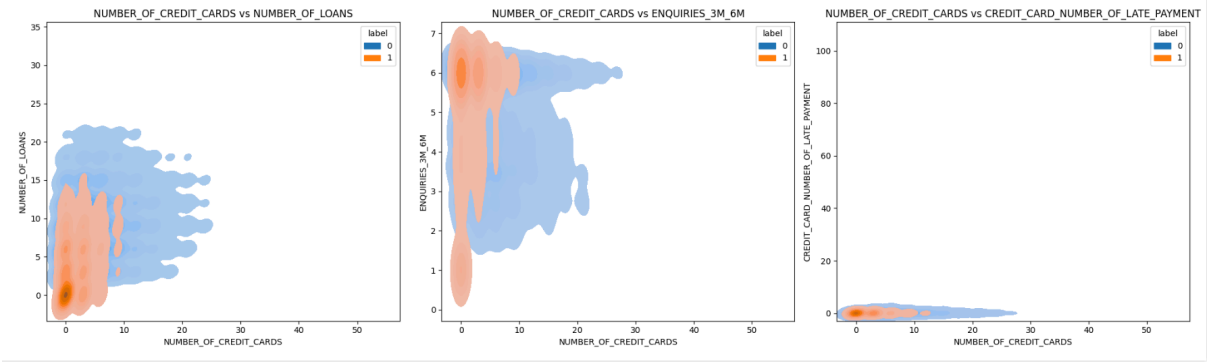
- Dữ liệu của đặc trưng: label 0 (81,78%) và label 1 chiếm (18.22%), điều đó cho thấy là dữ liệu của chúng ta có sự mất cân bằng, do đó cần sử dụng kĩ thuật resampling để cân bằng bộ dữ liệu.



- SHORT_TERM_COUNT chiếm đa số cho thấy khách hàng ưa chuộng vay ngắn hạn. Tuy nhiên, tỉ lệ về vay ngắn, trung và dài hạn trên cả label 0 và 1 khá tương tự nhau. Vì vậy, không thể căn cứ vào số khoản nợ để đánh giá một khách hàng có khả năng trả nợ đúng hạn hay không.



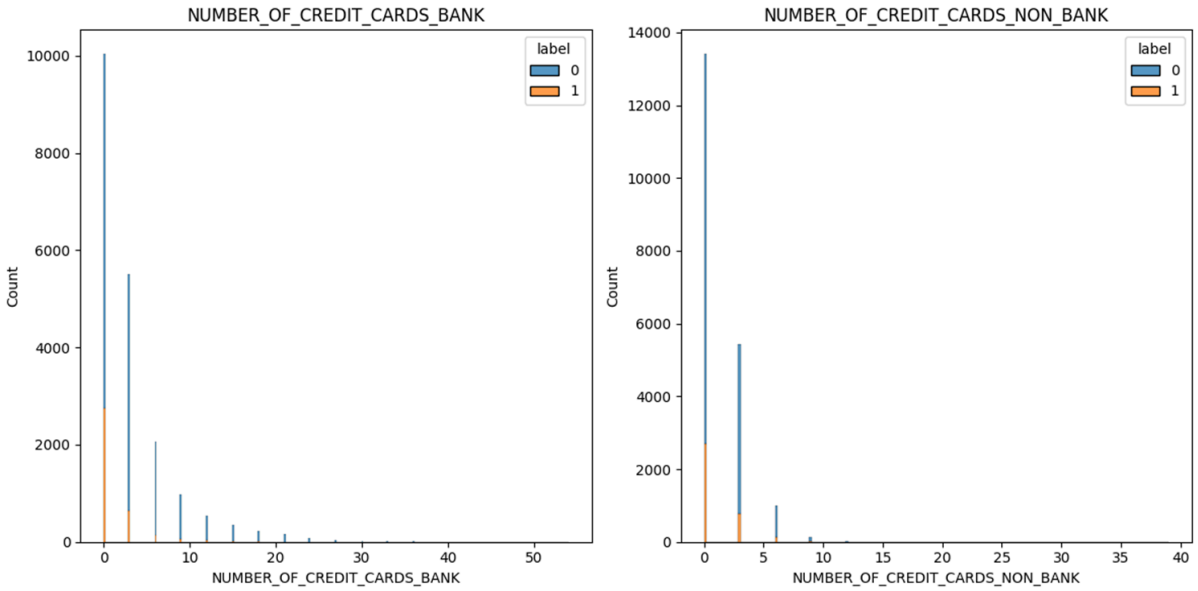
- Dễ thấy rằng số lượt tra cứu tín dụng liên quan đến sản phẩm tài chính càng ít thì khả năng khách hàng trả nợ không đúng hạn càng cao và ngược lại.
- Tuy nhiên, thời điểm thực hiện tra cứu tín dụng không ảnh hưởng tới khả năng trả nợ của khách hàng.



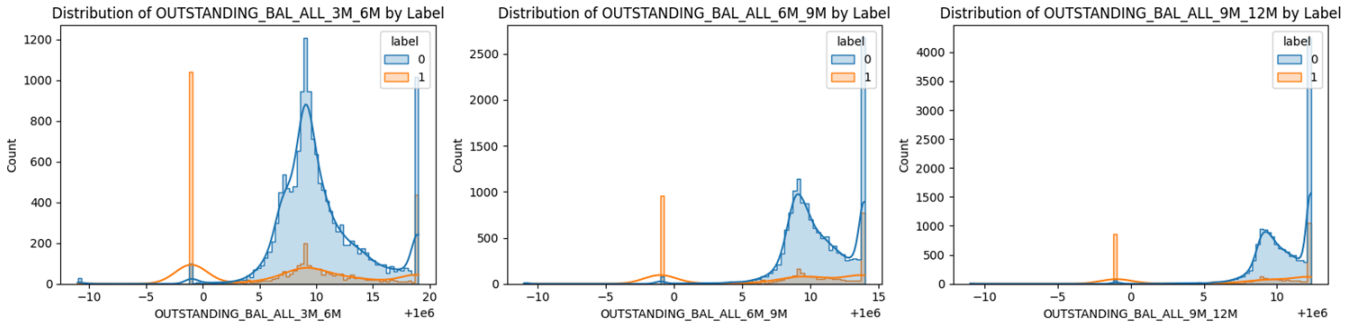
- Nhận thấy rằng những khách hàng không có khả năng trả nợ đúng hạn thường có ít thẻ tín dụng được cấp từ ngân hàng hay tổ chức phi ngân hàng.

- Đặc biệt, khách hàng vừa có ít thẻ tín dụng, vừa có ít lượt tra cứu (1-2 lượt) hoặc không tra cứu tín dụng liên quan đến các sản phẩm tài chính trong 3-6 tháng gần đây có khả năng rất cao sẽ trả nợ không đúng hạn.

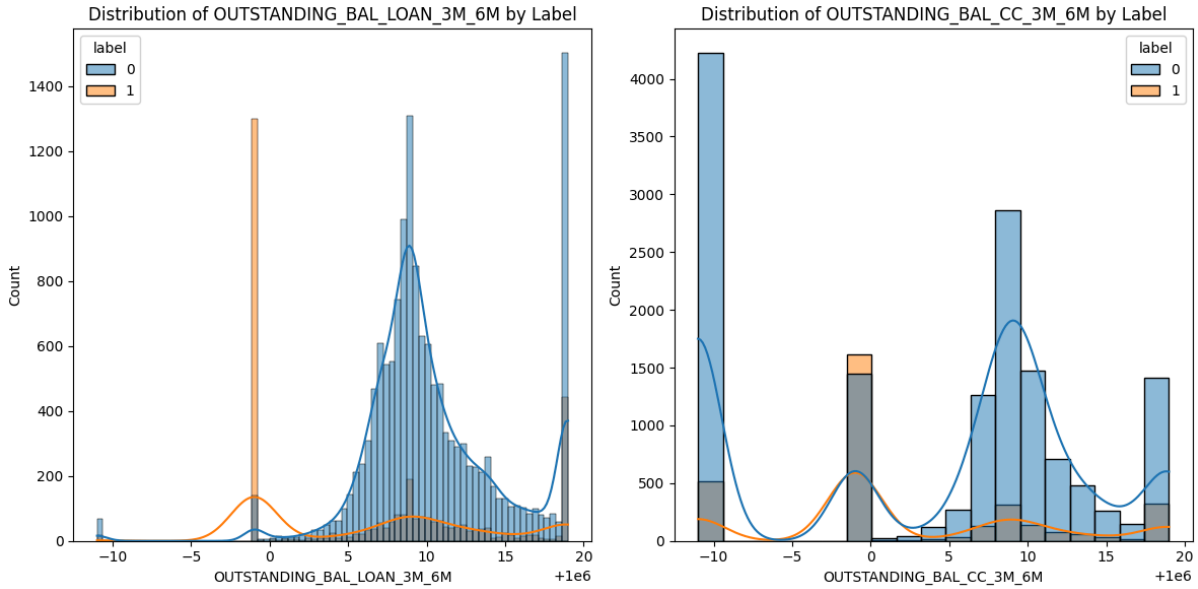
- Ngoài ra, cũng khá bất ngờ khi những khách hàng không có khả năng trả nợ đúng hạn có ít lần thanh toán trễ trên các thẻ tín dụng. Điều này có thể do khách hàng thuộc nhóm này có số lượng thẻ tín dụng khá ít.



- Như đã thấy, khách hàng không có khả năng trả nợ đúng hạn sẽ có xu hướng sử dụng ít thẻ tín dụng. Tuy nhiên, việc thẻ tín dụng được cấp bởi ngân hàng hay tổ chức phi ngân hàng không ảnh hưởng tới việc xác định khả năng trả nợ của khách hàng.



- Số dư nợ cao dường như có dấu hiệu tốt cho thấy khả năng trả nợ đúng hạn, điều này có thể được giải thích do thường nhóm khách hàng này có xu hướng quản lý tài chính tốt. Còn khách hàng có số dư nợ nhỏ dễ rơi vào nhóm trả nợ quá hạn. Đặc biệt, qua thời gian, sự phân biệt lại càng ngày càng rõ ràng hơn nữa đặc biệt là vào thời điểm 9-12 tháng, sự chênh lệch rất rõ rệt.
-



3.2 Kết luận:

- Mất cân bằng dữ liệu**:** Label 0 (81.78%) và label 1 (18.22%) cho thấy dữ liệu không cân bằng.
- Vay ngắn hạn phổ biến: Tỉ lệ vay ngắn, trung, và dài hạn trên cả hai label khá tương tự, không đủ để đánh giá khả năng trả nợ.
- Lượt tra cứu tín dụng: Số lượt tra cứu tín dụng liên quan đến sản phẩm tài chính càng ít thì khả năng khách hàng trả nợ không đúng hạn càng cao và ngược lại. Việc khách hàng có trả nợ đúng hạn hay không không phụ thuộc vào thời điểm tra cứu tín dụng.
- Số lượng thẻ tín dụng: Khách hàng không trả nợ đúng hạn có ít thẻ tín dụng, nhưng số lượng thẻ từ ngân hàng hay tổ chức phi ngân hàng không ảnh hưởng nhiều đến khả năng trả nợ.
- Số dư nợ: Dễ hiểu khi khách hàng không có khả năng trả nợ đúng hạn thường có số dư nợ tăng lên theo thời gian và ngược lại.

=> Nhóm khách hàng có khả năng cao trả nợ không đúng hạn là :

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

=> Nhóm khách hàng có khả năng trả nợ đúng hạn có một trong những đặc trưng sau đây:

- Nhiều lượt tra cứu tín dụng (>2 lượt)
- Nhiều thẻ tín dụng (>=10)
- Số dư nợ giảm dần

Ví dụ: Ta so sánh 2 customer với id 1639,23717. Xét số lượt tra cứu tín dụng 3 tháng gần đây, id 1639 có 37 lượt nhưng ID 23717 có 7 lượt, tuy nhiên lại có 72 lượt với 12 tháng trở lại đây trong khi ID 1639 chỉ có 17 lượt. Các xu hướng cũng được chứng minh ở Outstanding và number of credit cards.

customer_id	ENQUIRIES_3M	ENQUIRIES_12M	OUTSTANDING_BAL_LOAN_CURRENT	NUMBER_OF_CREDIT_CARDS_BANK
1639	37.0	17.0	1000000.0	1.0
23717	7.0	72.0	1001670.0	7.0

4. Mô hình dự đoán

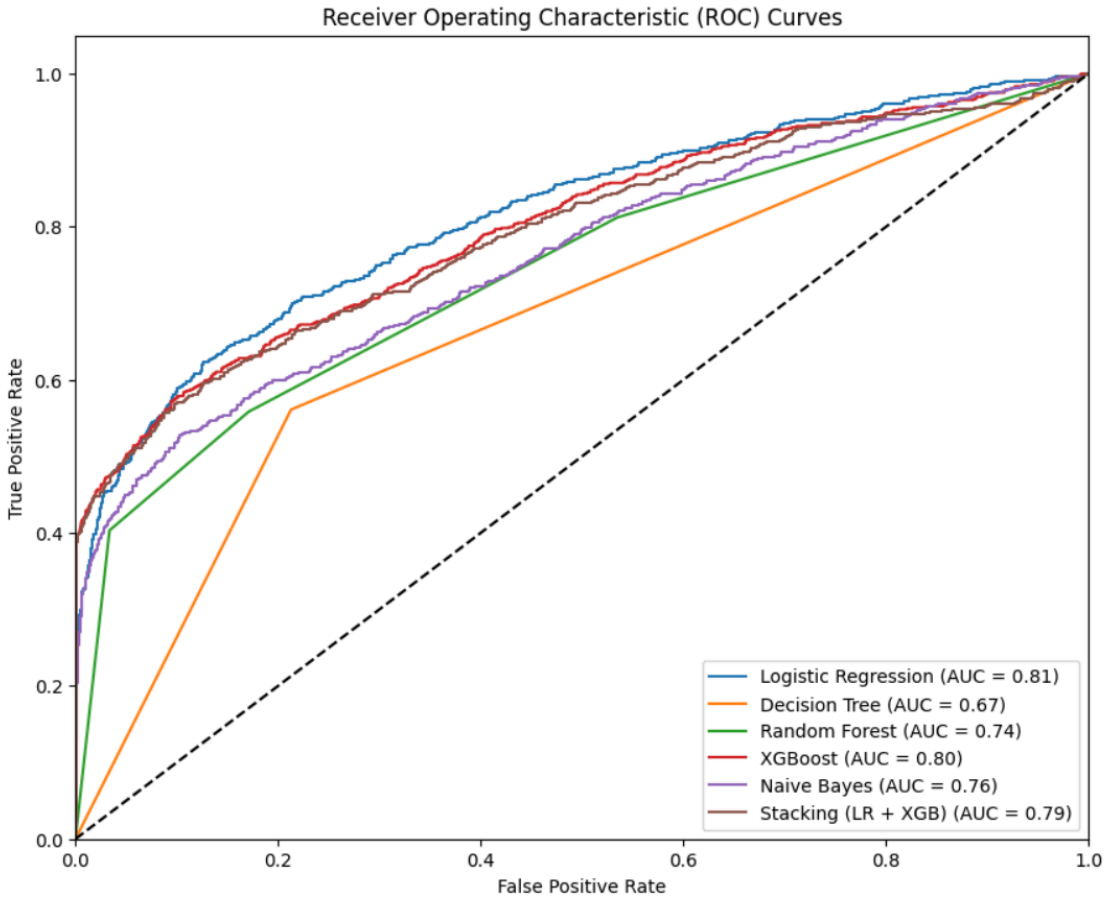
4.1. Xây dựng giải pháp:

- Đội thi sau quá trình tìm hiểu đã đưa ra giải pháp sử dụng mô hình học máy để dự báo khả năng thanh toán của khách hàng dựa trên các dữ liệu đã trực quan hóa.
- Sau khi chọn được 30PCs, ta sẽ đem chúng để huấn luyện. Ở đây ta sẽ thử nghiệm ở nhiều mô hình học máy khác nhau.
- Sau khi tiền xử lý và trích chọn dữ liệu, đội thi tiến hành chia dữ liệu thành 80% dùng cho việc huấn luyện và 20% dùng cho việc đánh giá mô hình (dữ liệu được chia theo đúng tỉ lệ đối với cả giá trị 0 và 1 của '*label*').
- Sau đó, nhận thấy là có sự chênh lệch giữa label 0 và 1 nên chúng ta sẽ resampling để cân bằng lại dữ liệu. Ở đây, đội thi dùng SMOTE.

4.2. Lựa chọn mô hình

- Đánh giá hiệu suất các 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



- Các chỉ số thực tế: AUC, Gini, Kolmogorov-Smirnov.
    + AUC: là một chỉ số đo lường hiệu suất của mô hình phân loại. Ở đó, AUC là diện tích dưới đường cong ROC (Là đồ thị thể hiện mối quan hệ giữa TPR và FPR). Giá trị AUC dao động từ 0 đến 1 và cho biết khả năng phân biệt giữa hai lớp của mô hình. Càng gần 1 thì mô hình càng hoàn hảo, và 0.5 cho thấy mô hình không khác gì phân loại ngẫu nhiên.

    + GINI:
        - GINI = 2 * AUC - 1, là một chỉ số chuyên dùng để tính rủi ro tín dụng, trong đó thì:
        - GINI = 1: Mô hình hoàn hảo, phân biệt chính xác tất cả các trường hợp.
        - GINI = 0: Mô hình không phân biệt được tốt hơn việc đoán ngẫu nhiên.
        - GINI < 0: Mô hình phân loại kém, sai lệch hoàn toàn với thực tế.

- GINI > 0.6: Mô hình hoạt động tốt

+ Kolmogorov-Smimov (SV): là giá trị lớn nhất của sự khác biệt tuyệt đối giữa hai hàm phân phối tích lũy (CDF) của hai nhóm dữ liệu (giữa có khả năng trả nợ đúng hạn và không có khả năng trả nợ đúng hạn). Giá trị KS dao động từ 0 đến 1:
  - KS = 0: Mô hình không có khả năng phân biệt giữa hai nhóm.
  - KS > 0.4: Mô hình có khả năng áp dụng tốt trong thực tế.
  - KS = 1: Mô hình hoàn hảo trong việc phân biệt giữa hai nhóm.

- Mô hình Logistic Regression đưa ra các giá trị tốt nhất ở cả 3 chỉ số trong 6 mô hình đang xét.

4.3. Đánh giá giải pháp

- Ưu điểm:
  + Mô hình cho ra AUC là 0.81, Gini là 0.63 và KS là 0.50 cho thấy mô hình phân biệt 2 nhóm khách hàng rất tốt.
  + Mô hình được lựa chọn là Logistic Regression, một mô hình dễ hiểu và dễ triển khai trong doanh nghiệp, cùng với đó là tính ổn định cao.
- Nhược điểm:
  + Do Logistic Regression là một model khá đơn giản vậy nên nó sẽ bị kém hiệu quả đối với các đặc trưng đa cộng tuyến.

  => Giải pháp ở đây là sử dụng PCA để giảm chiều dữ liệu nhằm tránh đa cộng tuyến giữa các đặc trưng.

5. Ứng dụng thực tiễn

- Scorecard là một công cụ đánh giá được sử dụng để phân tích khả năng tín dụng của một cá nhân hoặc doanh nghiệp. Nó giúp các tổ chức tài chính xác định rủi ro liên quan đến việc cấp tín dụng.

- Cách áp dụng scorecard :
  Ví dụ: Ta sẽ tính trên hai người A và B, với đặc điểm là : người A :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 ; người B là cập vào tín dụng trong 3 tháng gần đây 7 lần, 12 tháng là 72 lần, Tổng số dư nợ trên mọi nền tảng là 100tr, Có 7 thẻ tín dụng.

Ta sẽ quy ước mức điểm 500 có tỷ lệ odds là 1 phần 50 và người xây dựng mô hình muốn tỷ lệ odds tăng gấp đôi sau mỗi lần giảm 20 điểm (tức là pdo bằng -20), thì scorecard cho A và B sẽ là 411.36 và 362.5 trong khi tiêu chuẩn ở đây là 386 => A dự đoán không có khả năng trả nợ còn B có khả năng trả nợ.

- Với ứng dụng trong thực tiễn, chúng ta sẽ xem xét một số cách triển khai mô hình dự đoán này cho doanh nghiệp. Với một mô hình, ta sẽ có những bước như sau: thứ nhất là thu thập và khám phá dữ liệu (EDA). Thì ở đây chúng ta sẽ thu thập dữ liệu từ những lần giao dịch của khách hàng, tiền xử lý, làm sạch nó và trực quan hoá chúng để hiểu rõ về thông tin dữ liệu.

- Tiếp theo, chúng ta sẽ dùng dữ liệu được xử lý đó, chia thành 2 tập train với test và huấn luyện và huấn luyện nó, sau đó đánh giá hiệu suất. Và cuối cùng, mô hình sau khi được huấn luyện xong sẽ được tích hợp vào trang web hay phần mềm quản lý khách hàng của các doanh nghiệp.

- Sau khi tích hợp vào phần mềm, các doanh nghiệp sẽ dựa vào đó, sử dụng Log-odds cho mô hình, sau đó tinh chỉnh bằng Offset và Factor để tính Scorecard, từ đó dựa vào điểm số đó phân loại các cụm khách hàng.

- Ở đó, với dự báo khả năng trả nợ đúng hạn thì doanh nghiệp sẽ dễ dàng duyệt vay với khách hàng có khả năng trả nợ, ngược lại thì những người không có khả năng sẽ áp dụng các điều khoản và chính sách khắt khe hơn để tránh rủi ro tài chính (có tính chất dây chuyền nên khá nguy hiểm), và đặc biệt là kiểm soát tỉ lệ nợ xấu ở ngưỡng an toàn và ổn định. Không chỉ vậy với mô hình này doanh nghiệp có thể cá nhân hoá dịch vụ của khách hàng, và sẽ đảm bảo quyền lợi giữa đôi bên.

- Tuy nhiên, có một số lưu ý là để mô hình giữ được chính xác nhất thì ta nên cập nhật data thường xuyên và có thể kiểm tra hiệu suất định kì, tránh khả năng underfitting hay overfitting. Không chỉ vậy, thậm chí chúng ta có thể đổi biến để có thể cải thiện hiệu suất mô hình.