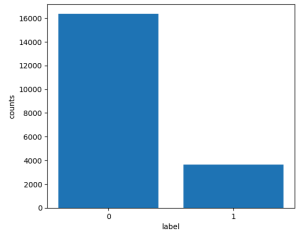
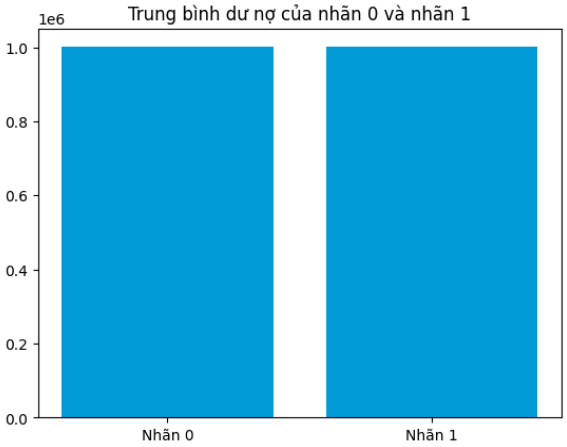
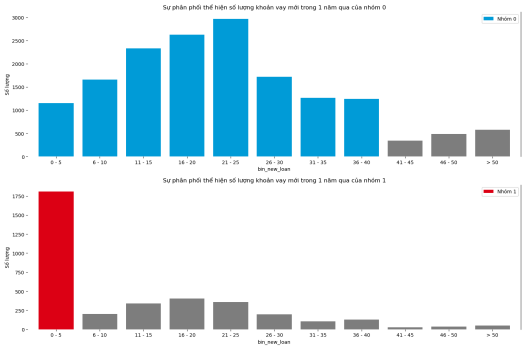
Ảnh 1: Tổng số người 0 áp đảo người 1



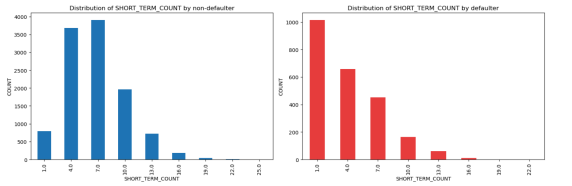
Ảnh 2: Trung bình chỉ số giữa 0 và 1



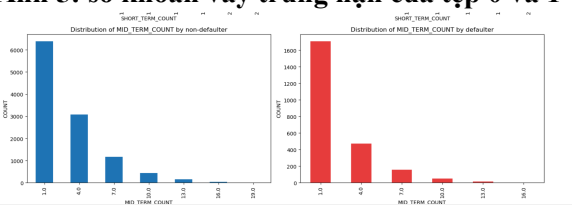
Ảnh 3: Số khoản vay mới của tệp 0 và 1



Ảnh 4: Số khoản vay ngắn hạn của tệp 0 và 1



Ảnh 5: số khoản vay trung hạn của tệp 0 và 1



Ảnh 6: Phân phối các mối quan hệ của nhóm 0 và nhóm 1

REPORT DATA SCIENCE TALENT COMPETITION 2024

1. DATA CLEANING

1.1 Dữ liệu bị thiếu

- Các giá trị bị thiếu trải dài ở hầu hết tất cả các dòng, vì thế để thuận tiện cho EDA, missing value sẽ được thay thế bằng giá trị trung bình

1.2 Dữ liệu lặp lại

- Bỏ Trường số ngày trễ nợ, số tăng lên trong dư nợ vì giá trị 431 và giá trị null xuất hiện trong 98% tệp dữ liệu => gần như 100% trường đó mang 1 giá trị

1.3 Sự mất cân bằng trong dữ liệu

- Có một số trường có một giá trị chiếm đại đa số (> 97%)
  => Loại các trường này

2. EDA: INITIAL FINDINGS/ OVERVIEW

Tìm điểm khác nhau giữa nhóm 0 và nhóm 1 bằng tay

2.1 Tệp khách hàng 0 hiện đang áp đảo so với tệp 1

Số lượng khách hàng trong tập 0 **nhiều gấp 4 lần** tập 1. Vì vậy, tổng các dữ liệu của tập 0, chẳng hạn như số khoản vay, số lần tra cứu, số quan hệ, sẽ luôn có **sự chênh lệch lớn hơn nhiều** so với tập 1. Tuy nhiên, nếu xét về **giá trị trung bình** của các chỉ số như số tiền vay nợ, số lần quan hệ, thì hai tập dữ liệu này lại **tương đương nhau**. Do đó, việc so sánh tổng số lượng giữa hai tập sẽ khiến tập 0 luôn có vẻ vượt trội hơn tập 1.

Để có được cái nhìn khách quan và chính xác, chúng ta nên **so sánh phân phối** của tập 0 và tập 1 một cách riêng biệt.

2.3 Sự khác nhau giữa Tệp 1 và tệp 0

a/ Số khoản vay mới:

Sự khác biệt về khả năng tiếp cận khoản vay mới giữa hai nhóm khách hàng được thể hiện rõ ràng trong hình 3. Trong khi nhóm đúng hạn (tệp 0) có số khoản vay phân bố đều từ 0 đến 40, thì nhóm quá hạn (tệp 1) lại chủ yếu tập trung ở khoảng 0-5. Rõ ràng, **lịch sử tín dụng kém** đang ảnh hưởng đáng kể đến **khả năng vay vốn** của nhóm quá hạn.

b/ Số khoản vay ngắn hạn/ dài hạn/ trung hạn:

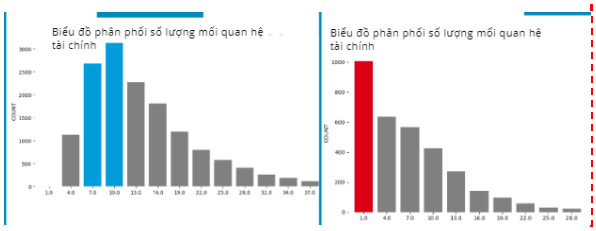
Trung bình số khoản vay ngắn hạn của nhóm khách hàng đúng hạn (tệp 0) tập trung vào khoảng 4-10 khoản, nhiều hơn đáng kể so với nhóm khách hàng quá hạn (tệp 1) - thường chỉ vay 1-4 khoản. Mặc dù vậy, số khoản vay trung hạn của hai nhóm là ngang nhau. Có thể thấy rằng việc vay nhiều khoản ngắn hạn có liên quan đến việc duy trì trạng thái đúng hạn.

c/ Quan hệ với tổ chức tài chính/ phi tài chính:

Nhóm khách hàng đúng hạn (tệp 0) thường có quan hệ với ít nhất 7 tổ chức tài chính, trong khi nhóm quá hạn (tệp 1) thì không. Đáng chú ý là nhóm quá hạn lại có xu hướng quan hệ nhiều hơn với các tổ chức phi tài chính so với tổ chức tài chính.

Điều này cho thấy nhóm khách hàng quá hạn có thể gặp khó khăn trong **việc tiếp cận các tổ chức tài chính truyền thống,** dẫn đến việc phải **tìm đến các tổ chức phi tài chính**

Tóm tắt sự khác biệt giữa nhóm 0 và nhóm 1:



Ảnh 7: Các biến mới được tạo sau bước feature engineer

Feature 1: Binary Features (Short Term, Mid Term, Long Term)

```
df_t = df[['SHORT_TERM_COUNT', 'MID_TERM_COUNT', 'LONG_TERM_COUNT', 'label']].copy(deep=True)
df_t['MANY_SHORT'] = np.where(df_t['SHORT_TERM_COUNT'] >= 4, 1, 0)
df_t['MANY_MID'] = np.where(df_t['MID_TERM_COUNT'] >= 4, 1, 0)
df_t['MANY_LONG'] = np.where(df_t['LONG_TERM_COUNT'] >= 4, 1, 0)
df_t['Theory 1'] = (df_t['MANY_SHORT'] * 10 + df_t['MANY_MID'] * 5 + df_t['MANY_LONG']) / 15
df['Diff_in_term_count'] = df_t['Theory 1']
```

Feature 2: Binary Features (Number of Loans, Number of Loans Bank, Number of Loans Non Bank)

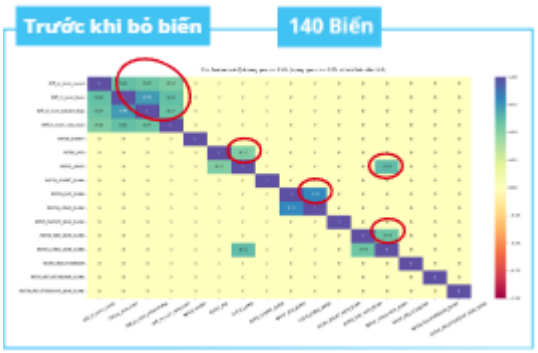
```
df_t = df[['NUMBER_OF_LOANS', 'NUMBER_OF_LOANS_BANK', 'NUMBER_OF_LOANS_NON_BANK', 'label']].copy
(deep=True)
df_t['MANY_LOANS'] = np.where(df_t['NUMBER_OF_LOANS'] >= 4, 1, 0)
df_t['MANY_LOANS_BANK'] = np.where(df_t['NUMBER_OF_LOANS_BANK'] >= 4, 1, 0)
df_t['MANY_LOANS_NON_BANK'] = np.where(df_t['NUMBER_OF_LOANS_NON_BANK'] >= 4, 1, 0)
df_t['Theory 1'] = df_t.iloc[:, -3:].sum(axis=1)
df['Diff_in_num_loan'] = df_t['Theory 1']
```

Feature 3: Differece in 3 features realted to Number of Relationships ¶

```
df_t = df[['NUMBER_OF_RELATIONSHIP', 'NUMBER_OF_RELATIONSHIP_BANK', 'NUMBER_OF_RELATIONSHIP_NON_B
ANK', 'label']].copy(deep=True)
# df_t.median()
df_t['MANY_RELATIONSHIP'] = np.where(df_t['NUMBER_OF_RELATIONSHIP'] >= 4, 1, 0)
df_t['MANY_RELATIONSHIP_BANK'] = np.where(df_t['NUMBER_OF_RELATIONSHIP_BANK'] >= 4, 1, 0)
df_t['MANY_RELATIONSHIP_NON_BANK'] = np.where(df_t['NUMBER_OF_RELATIONSHIP_NON_BANK'] >= 4, 1, 0)

df_t['Theory 1'] = (df_t['MANY_RELATIONSHIP'] * 5 + df_t['MANY_RELATIONSHIP_BANK'] + df_t['MANY_R
ELATIONSHIP_NON_BANK']) / 7
df['Diff_in_num_relationship'] = df_t['Theory 1']
```

Ảnh 8a (minh họa): Tương quan trước khi bỏ biến



Ảnh 8b(minh họa): Tương quan sau khi bỏ biến

- **Số lượng khoản vay theo thời gian**: Nhóm 0 luôn có trung bình cao hơn nhóm 1 (0% - 40%). Nhóm 0 tập trung chủ yếu vào 4-7 khoản vay ngắn hạn, trong khi nhóm 1 chỉ có 1 khoản vay.
- **Tổng số khoản vay, thẻ tín dụng, mối quan hệ tài chính, khoản vay mới theo tháng**: Trung bình nhóm 0 luôn cao hơn nhóm 1 (20% - 50%). Phân bố:
  - **Tổng số khoản vay**: Nhóm 0 tập trung vào 4-7 khoản vay, nhóm 1 chỉ có 1 khoản.
  - **Thẻ tín dụng**: Cả hai nhóm tương đồng, tập trung vào 1-4 thẻ.
  - **Mối quan hệ tài chính**: Nhóm 0 có 4-10 mối quan hệ, nhóm 1 chỉ có 1-4.
  - **Khoản vay mới**: Nhóm 0 tập trung vào 1 và 4 khoản vay mới trong 3 tháng trước, 4-7 khoản vay trong 6-12 tháng trước. Nhóm 1 chủ yếu có 1 khoản vay mới trong cả giai đoạn.

Tìm điểm khác nhau giữa nhóm 0 và nhóm 1 bằng máy

2.5 Feature Engineering: Tạo ra các nhóm feature mới nhằm có thêm ý nghĩa input vào máy

Nhóm Feature 1: Short Term, Mid Term, Long Term

- **Biến mới**: Tạo ra 3 biến nhị phân (MANY_SHORT, MANY_MID, MANY_LONG) đại diện cho số lượng khoản vay ngắn hạn, trung hạn, và dài hạn có giá trị lớn hơn hoặc bằng 4, sau đó, biến DIFF_IN_TERM_COUNT được tạo ra bằng cách kết hợp 3 biến nhị phân này.
- **Tác dụng đến model**: Việc nhóm các khoản vay ngắn, trung, và dài hạn giúp mô hình dễ dàng phân loại người vay có nhiều khoản vay ở các kỳ hạn khác nhau, giúp mô hình dự đoán tốt hơn về khả năng thanh toán của họ.

Nhóm Feature 2: Number of Loans, Number of Loans Bank, Number of Loans Non Bank

- **Biến mới**: Tạo ra các biến nhị phân MANY_LOANS, MANY_LOANS_BANK, và MANY_LOANS_NON_BANK để đánh dấu số lượng khoản vay tổng, từ ngân hàng, và từ các tổ chức không phải ngân hàng có giá trị lớn hơn một ngưỡng nhất định.Sau đó, biến DIFF_IN_NUM_LOAN được tạo ra bằng cách kết hợp 3 biến nhị phân này.
- **Tác dụng đến model**: Các biến này giúp mô hình nhận biết mức độ phụ thuộc của người vay vào các nguồn tài chính khác nhau (ngân hàng hay tổ chức phi ngân hàng), có thể ảnh hưởng đến khả năng rủi ro tín dụng

Nhóm Feature 3: Number of Relationships

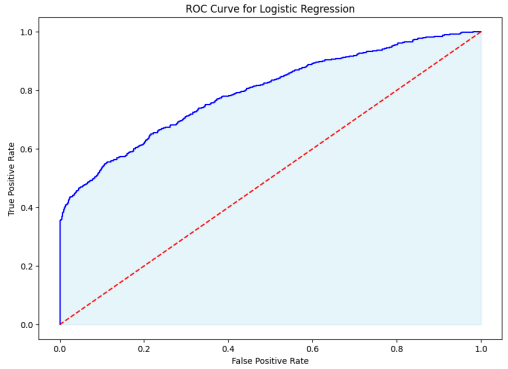
Biến mới: Tạo các biến nhị phân đánh dấu số lượng quan hệ tín dụng tổng, với ngân hàng, và với tổ chức phi ngân hàng nếu số quan hệ lớn hơn hoặc bằng 4. Sau đó, biến DIFF_IN_NUM_RELATIONSHIPS được tạo ra bằng cách kết hợp 3 biến nhị phân này.

Tác dụng đến model: Những biến này giúp mô hình xác định mức độ tương tác tín dụng của khách hàng với các tổ chức tài chính, từ đó có thể

Ảnh 9: IV và feature importance

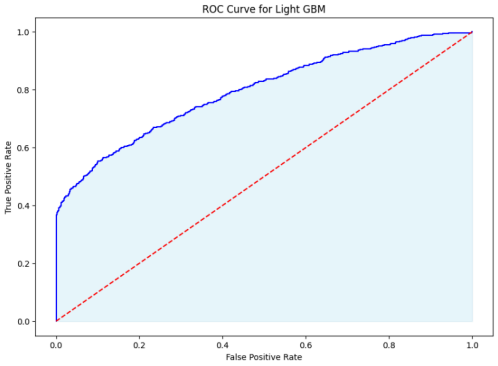
	column	IV	rank
124	RATIO_LONG	1.139427	suspicious
43	OUTSTANDING_BAL_LOAN_3M_6M	1.095279	suspicious
32	OUTSTANDING_BAL_LOAN_6M	1.021502	suspicious
31	OUTSTANDING_BAL_LOAN_3M	1.008555	suspicious
30	OUTSTANDING_BAL_LOAN_CURRENT	0.932673	suspicious
...	...	...	...
100	ENQUIRIES_FROM_BANK_FOR_CC_12M	0.000000	Useless
101	ENQUIRIES_FROM_NON_BANK_FOR_CC_12M	0.000000	Useless
102	ENQUIRIES_3M_6M	0.000000	Useless
103	ENQUIRIES_6M_9M	0.000000	Useless
104	ENQUIRIES_9M_12M	0.000000	Useless

Ảnh 10: Điểm ROC dùng Logistic Regression Baseline trong tập Val



Ảnh 11: Điểm ROC dùng Logistic Regression trong tập Val

Điểm ROC dùng LightGBM trong tập Val



Điểm ROC dùng Random Forest trong tập Val

đánh giá khả năng tín dụng của họ dựa trên mức độ tin tưởng từ nhiều nguồn tài chính.

Tương tự nhóm Feature 4 và 5 lần lượt có tác động như sau:

- Các biến mới được tạo ra trong quá trình feature engineering từ các biến nhị phân như MANY_NEW_LOAN_3M, MANY_NEW_LOAN_6M, v.v., đóng vai trò xác định **có nhiều khoản vay mới** trong các khoảng thời gian khác nhau (3 tháng, 6 tháng, 9 tháng, 12 tháng) dựa trên ngưỡng cụ thể (ví dụ, 4 khoản vay). Các biến này có thể được dùng để **đánh giá tần suất vay mới** của khách hàng trong một khoảng thời gian nhất định, hỗ trợ việc phân loại hoặc dự đoán rủi ro tín dụng.
- Các biến mới như RATIO_SHORT, RATIO_MID, RATIO_LONG, v.v., được tạo ra để tính tỷ lệ số lượng các khoản vay ngắn hạn, trung hạn, dài hạn, và mối quan hệ tín dụng so với tổng số khoản vay. Chúng giúp cung cấp thông tin về cơ cấu phân bổ của các loại khoản vay mà khách hàng đang có, từ đó **hỗ trợ đánh giá mức độ phụ thuộc của khách hàng** vào các loại hình vay khác nhau và tiềm năng rủi ro tài chính.

2.6 Feature selection

- Đây là 1 bộ dataset lớn, và vì thời gian từ khâu xử lý dữ liệu cho tới đưa vào mô hình phải rất nhanh chóng để có thể đáp ứng nhu cầu của khách hàng, thường là dưới 2 phút. Do đó, chỉ nên dùng một số lượng biến nhất định để có thể tối ưu thời gian xử lý của sản phẩm.

- Trước tiên là phân tích tương quan giữa các biến, với ngưỡng đặt ra là 50%, chúng tôi sẽ loại bỏ 1 trong 2 biến có độ tương quan lớn hơn 50%, với hình 10a là số lượng biến khi chưa loại bỏ, 140 biến, sau khi loại bỏ (hình 10b), còn **32 biến**. Việc này ngoài giúp giảm thời gian xử lý cũng sẽ đồng thời và giúp những phòng ban khác dễ dàng thông dịch.

- Ngoài ra, một kỹ thuật chọn biến được gọi là Weight Of Evidence (WOE), phương pháp này sẽ đóng thùng các giá trị liên tục và đánh giá mối quan hệ giữa chúng và kết quả mục tiêu. Sau khi có được WOE, sẽ ra được 1 giá trị gọi là Information Value (IV), giá trị này dùng để đánh giá khả năng dự đoán của một biến. Nếu IV càng cao, biến đó càng quan trọng trong việc dự đoán kết quả. Nếu IV bé hơn 0.02, biến đó sẽ không có khả năng dự đoán, và ngược lại, biến đó sẽ có khả năng dự đoán tùy theo độ lớn của IV. Sau khi sử dụng phương pháp này để đánh giá, số lượng biến còn lại là **25 biến.**

3. MODELS:

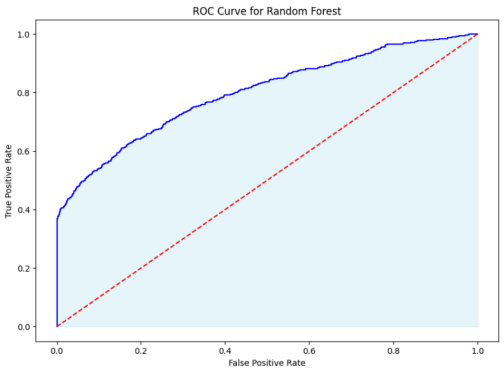
3.1/ Logistic Regression

- Mô hình tuyến tính (linear), dựa trên xác suất để dự đoán kết quả nhị phân.
- Đây là một mô hình đơn giản, dễ hiểu, giải thích rõ ràng mối quan hệ giữa các biến độc lập và xác suất vỡ nợ. Nó thích hợp khi dữ liệu tuyến tính và giúp giải thích dễ dàng kết quả.

*Logistic Regression Baseline: Kiểm thử model trước với tập validation trên data gốc (sau khi fillna, chưa feature engineer) để biết được điểm Gini và AUC baseline

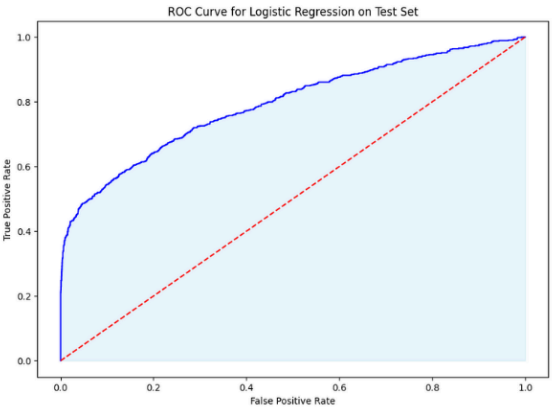
3.2/ Decision Tree:

- Mô hình tree-based, tạo ra các cây quyết định bằng cách chia nhỏ dữ liệu dựa trên các điều kiện.

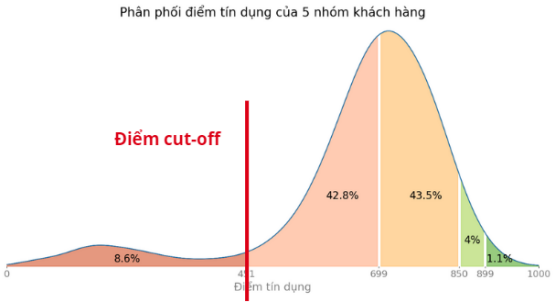


Ảnh 12: Kết quả mô hình

MÔ HÌNH	ĐIỂM GINI (2*AUC SCORE - 1)	ĐIỂM AUC
Logistic Regression	**0.563**	**0.781**
Light GBM	0.559580	0.779790
XGBoost	0.534233	0.767117
Random Forest	0.483200	0.741600
Decision Tree	0.468734	0.734367
Baseline Logistic Regression	0.300519	0.650259



Ảnh 12: Điểm tín dụng phân loại tệp khách hàng



- Mô hình này trực quan và dễ diễn giải, cung cấp các quy tắc đơn giản để xác định các yếu tố gây rủi ro tín dụng. Nó cho phép phân chia dữ liệu thành các nhóm dựa trên các đặc điểm khác nhau.

3.3/ Random Forest:

- Random Forest giúp giảm overfitting và tăng độ chính xác bằng cách kết hợp nhiều cây quyết định. Nó hiệu quả trong việc xử lý dữ liệu không đồng nhất và giảm rủi ro sai lệch khi ra quyết định.
- Mô hình tree-based, kết hợp nhiều cây quyết định độc lập và lấy trung bình kết quả.

3.4/ LightGBM:

- Mô hình tree-based, sử dụng boosting, nhưng xây dựng cây quyết định từ lá thay vì từ gốc, giúp tối ưu tốc độ.
- LightGBM nhanh hơn và tối ưu hơn khi xử lý các tập dữ liệu lớn. LightGBM thích hợp cho phân tích rủi ro tín dụng khi có dữ liệu lớn và phức tạp.

3.5/ XGBoost:

- Mô hình tree-based, sử dụng boosting với tối ưu hóa tốc độ và hiệu suất thông qua việc xử lý tuần tự và điều chỉnh lỗi.
- XGBoost xử lý tốt dữ liệu không cân bằng và giảm thiểu lỗi bằng cách tối ưu hóa mạnh mẽ, phù hợp trong các trường hợp yêu cầu độ chính xác cao.

4. Kết quả model

- Logistic Regression dự đoán trên tập test có Điểm GINI là **0.58** , cao hơn so với tập validation là **~2%.**
- Điều này cho thấy mô hình có khả năng tổng quát hóa tốt, tức là mô hình không bị overfitting (quá khớp) trên tập train và vẫn hoạt động tốt trên tập dữ liệu mới (tập test).

5. Chấm điểm tín dụng

a) Tính ra điểm tín nhiệm (credit scorecard) của mỗi khách hàng bằng cách tính điểm số cho mỗi feature (feature ở đây là một khoảng bin của biến liên tục hoặc một class của biến category). Điểm sẽ được scale theo công thức sau:

$$Score = (\beta.WOE+\alpha.n).Factor+Offset.n$$

Trong đó:

- β: Hệ số của biến trong phương trình hồi qui logistic.
- α: Hệ số chặn của phương trình hồi qui logistic.
- WOE: Hệ số trọng số bằng chứng của mỗi feature.
- n: Số lượng các biến của mô hình.
- Factor,Offset: Là các tham số được thiết lập để tính Score.

- Sau đó điểm tín dụng cuối cùng sẽ được scale trên thang FICO là 1000 điểm cho tất cả các khách hàng sử dụng **MinMaxScaler()**

b) Phân chia tệp khách hàng

- Trục ngang (x) đại diện cho "Điểm tín dụng", với điểm thấp nhất là 0 và điểm cao nhất là 1000.
- Trục đứng (y) là tỷ lệ phân phối điểm tín dụng.

Phân loại tệp khách hàng					
Điểm tín dụng	0-451	452-699	700-850	851-899	900-1000
Tỉ lệ quá hạn	93%	17.1%	7.3%	3.8%	0%
Phân khúc khách hàng	Rủi ro cao	Rủi ro vừa	Rủi ro trung bình	Rủi ro thấp	Rủi ro rất thấp

- Các phần chính của biểu đồ:
+ **Màu đỏ** (từ 0 đến khoảng 441): Đại diện cho nhóm khách hàng có điểm tín dụng thấp, có rủi ro cao (8.6%).
+ **Màu cam** (khoảng 441 đến 699): Nhóm khách hàng với rủi ro trung bình (42.8%).
+ **Màu vàng** (khoảng 699 đến 850): Nhóm khách hàng có rủi ro trung bình thấp (43.5%).
+ **Màu xanh nhạt** (850 đến 899): Nhóm khách hàng ít rủi ro (4%).
+ **Màu xanh đậm (trên 899):** Nhóm khách hàng rất ít rủi ro (1.1%).

6. Ngưỡng cutoff

Ngưỡng cut off sẽ giao thoa ở điểm 451 (khoảng chia khách hàng rủi ro và khách hàng không) do:

- **Điểm chuyển tiếp rõ rệt về tỷ lệ quá hạn:** Nhìn vào biểu đồ phân phối điểm tín dụng, ta thấy tỷ lệ quá hạn giảm mạnh khi điểm tín dụng vượt qua ngưỡng 451. Cụ thể, nhóm khách hàng có điểm tín dụng từ 0-451 có tỷ lệ quá hạn lên đến 93%, trong khi nhóm 452-699 chỉ còn 17.1%. Điều này cho thấy ngưỡng 451 là một ranh giới quan trọng để phân biệt khách hàng rủi ro cao và rủi ro thấp hơn.

- **Cân bằng tương đối giữa số lượng khách hàng và tỷ lệ quá hạn:** Vị trí điểm 451 trên biểu đồ cho thấy sự cân bằng tương đối giữa số lượng khách hàng và tỷ lệ quá hạn. Nó nằm gần giao điểm của đường thể hiện số lượng khách hàng và đường thể hiện tỷ lệ quá hạn. Điều này có nghĩa là việc chọn ngưỡng 451 giúp phân loại khách hàng thành các nhóm có số lượng tương đối đồng đều, đồng thời vẫn đảm bảo kiểm soát rủi ro tín dụng.

- Ngưỡng điểm 451 phân loại khách hàng thành 3 nhóm rủi ro tín dụng: nhóm 0-451 có rủi ro CAO (93% quá hạn), nhóm 452-699 có rủi ro TRUNG BÌNH (17.1% quá hạn) và nhóm 700-850 có rủi ro THẤP (0% quá hạn)

- **Quản lý rủi ro:** Ngưỡng cutoff giúp ngân hàng hoặc tổ chức tín dụng xác định rõ ràng các nhóm khách hàng có mức độ rủi ro khác nhau, từ đó áp dụng các chính sách tín dụng phù hợp. Ví dụ, với nhóm khách hàng có điểm tín dụng dưới 451 (rủi ro cao), có thể áp dụng hạn mức tín dụng thấp hơn, lãi suất cao hơn hoặc yêu cầu thêm tài sản đảm bảo.