

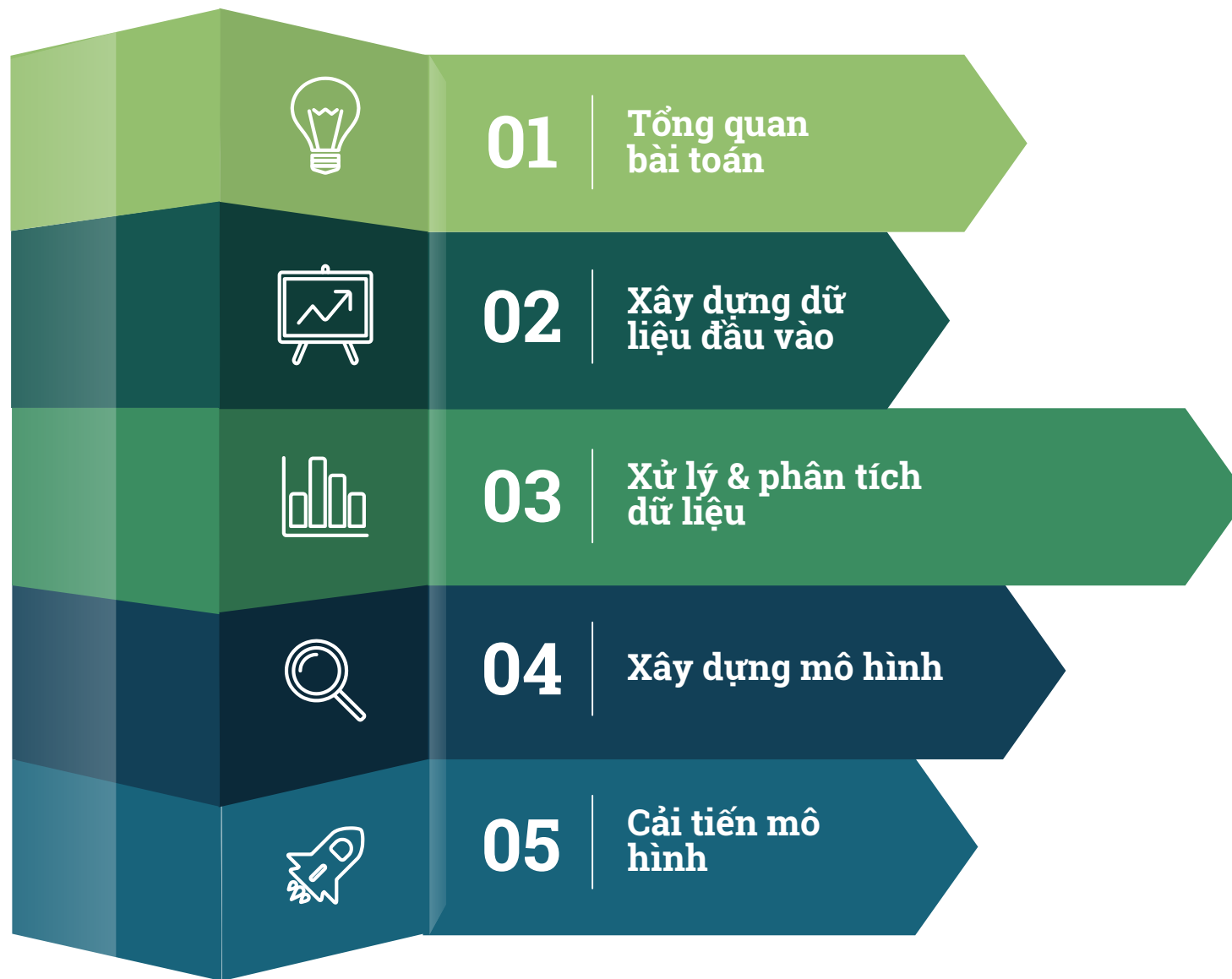


Báo cáo kỹ thuật dự án xếp hạng tín dụng

Phòng Phân tích dữ liệu – TTCN - VDS

Hà Nội, T5/2020

Nội dung báo cáo



1

Tổng quan bài toán

Introduction of model



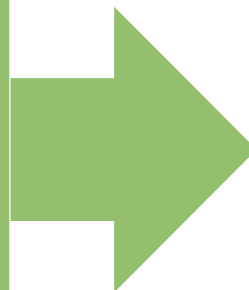
1. Mô hình hóa bài toán

Lợi thế xây dựng mô hình



80%

Khách hàng tới vay vốn chưa có lịch sử tín dụng gây khó khăn cho việc chấm điểm tín dụng bằng phương pháp truyền thống*



Đã có nhiều tổ chức sử dụng dữ liệu viễn thông để xếp hạng



trusting social



kt



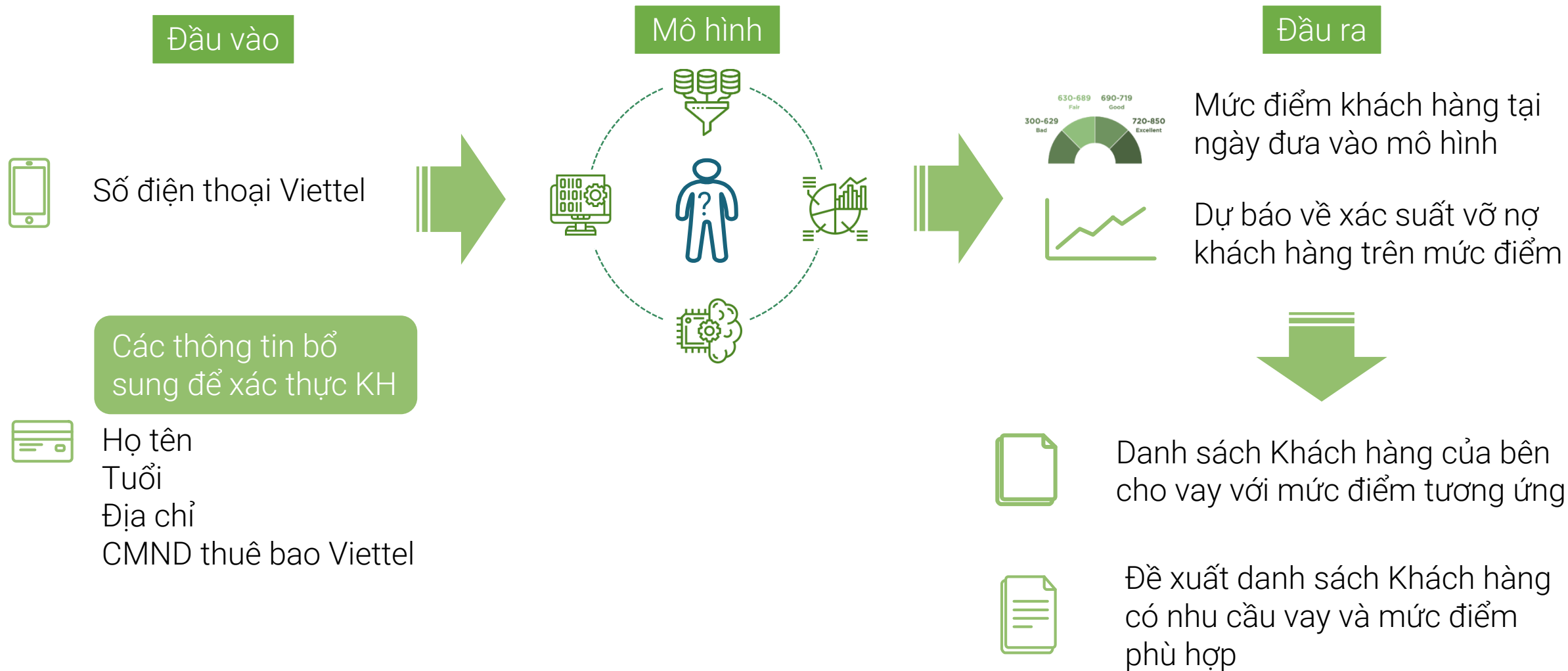
Viettel với lợi thế rất lớn khi chiếm 55% thị phần viễn thông, có khả năng giúp cho hàng triệu người Việt Nam có điểm tín dụng, để tiếp cận nguồn vốn, đồng thời, giúp các tổ chức tín dụng giảm thiểu rủi ro khi giải ngân.



1. Mô hình hóa bài toán

Yêu cầu bài toán

Xếp hạng tín dụng cho các khách hàng có nhu cầu vay tín chấp tiêu dùng. Mô hình minh họa:



2

Xây dựng dữ liệu đầu vào

Features Engineering



2. Lấy dữ liệu đầu vào

Mô tả dữ liệu nhãn



Khách hàng tốt (Good = 0)
Nợ quá hạn ≤ 30 ngày
Chiếm 96.92% bộ dữ liệu



Khách hàng xấu (Bad = 1)
Nợ quá hạn > 30 ngày
Chiếm 3.08% bộ dữ liệu

- Theo quy định Chính phủ, các mức nợ quá hạn trên 90 ngày là nợ dưới chuẩn, và thường được các định chế tài chính gán nhãn nợ xấu. Để phù hợp với mức chịu rủi ro của gói vay tiêu dùng, mô hình sẽ gán nợ xấu cho nợ quá hạn trên 30 ngày
- Nhãn nợ xấu được lấy từ sms banking, của các tổ chức tín dụng với sản phẩm vay tiêu dùng như

FE CREDIT
VAY TIÊU DÙNG TÍN CHẤP

SHB Finance
TÀI CHÍNH TIÊU DÙNG

HD SAISON
Tài chính tiêu dùng

CASH24

TPBank
Vì chúng tôi hiểu bạn

mcredit
Cần vay, được ngay

MIRAE ASSET
Finance Company

SHINHAN BANK
VIETNAM



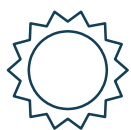
2. Lấy dữ liệu đầu vào

Mô tả dữ liệu đầu vào

Tổng số lượng biến đầu vào là **458 biến** và có cập nhật dữ liệu cho mô hình hàng tuần để đảm bảo cải thiện mô hình liên tục. Bài toán sử dụng **06 mốc** thời gian khác nhau:



Trong tuần (thứ 2 – thứ 6)



Làm việc (7h30 – 17h00)



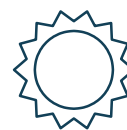
Nghỉ ngơi (17h00 – 23h00)



Buổi đêm (23h00 – 7h30)



Cuối tuần (thứ 7 – chủ nhật)



Làm việc (7h30 – 17h00)



Nghỉ ngơi (17h00 – 23h00)



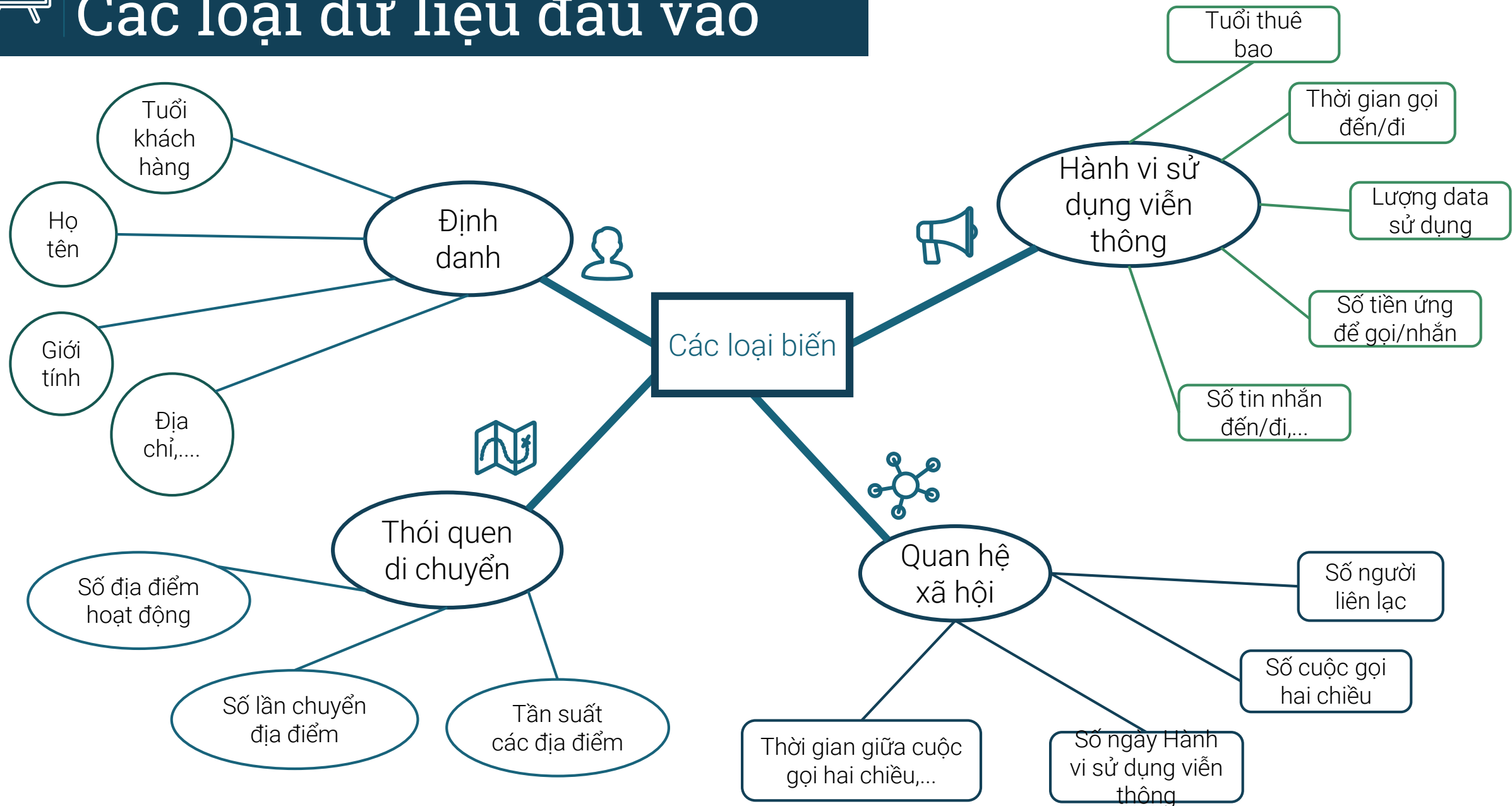
Buổi đêm (23h00 – 7h30)

- Tất cả các biến chứa dữ liệu viễn thông đều được tổng hợp theo 06 mốc này để phân biệt tốt nhất đặc điểm hành vi khách hàng.
- Dữ liệu được trích xuất trong **06 tháng liên tiếp**: Việc tăng thời gian quan sát nhằm tăng cường tính ổn định của những quan sát, và thấy rõ hơn xu hướng sử dụng thiết bị viễn thông của các thuê bao. Ngoài ra việc này cũng góp phần ngăn chặn một phần các sim rác mới hoạt động.



2. Lấy dữ liệu đầu vào






Các loại dữ liệu đầu vào





2. Lấy dữ liệu đầu vào

Dữ liệu về định danh

-  Tuổi khách hàng: Tuổi có phân bố tương quan với tỷ lệ nợ xấu (chi tiết trang 21)
-  Tuổi thuê bao: Thuê bao tuổi càng lớn thì càng đáng tin cậy, hạn chế khả năng sim rác
-  Họ và tên: Kiểm tra tính đầy đủ của thông tin thuê bao
-  Địa chỉ theo đăng ký của thuê bao: Tính chất vùng miền cũng có thể tác động lên tỷ lệ nợ xấu (xem trang 22)
-  Giới tính: Thường nữ giới có tỷ lệ nợ xấu nhỏ hơn nam giới

➤ Một thuê bao có thông tin định danh đầy đủ, rõ ràng, sẽ thường có độ tin cậy cao hơn



2. Lấy dữ liệu đầu vào

Dữ liệu hành vi sử dụng viễn thông



Lưu lượng data

Tổng lượng data đã sử dụng.
Tổng lượng data đã tải lên
Tổng lượng data đã tải xuống
Độ biến động data tải lên theo ngày.
Độ biến động data tải xuống theo ngày



Chi phí viễn thông

Tổng tiền đã sử dụng
Tổng tiền cho call
Tổng tiền cho sms
Tổng tiền cho data
VAS
Số lần ứng tiền
Tổng số tiền ứng
Trạng thái thuê bao: Trước, sau
Số lần thay đổi trạng thái thuê bao.



Đặc điểm nhắn/gọi

Tổng thời gian gọi đến/ gọi đi/ số sms,
số cuộc gọi nhớ
Độ biến động của các thời gian gọi đi,
gọi đến
Tần suất gọi/ nhắn tin
Số cuộc gọi nhớ
Số ngày thuê bao hoạt động viễn thông


➤ Các dữ liệu này thể hiện khả năng tài chính của thuê bao, mức độ chi trả của thuê bao cho các hoạt động viễn thông như một thuê bao chi trả đầy đủ và không ứng tiền điện thoại thường có xác suất vỡ nợ thấp hơn các thuê bao khác hay các thuê bao trả sau thường có mức độ tin cậy tốt hơn thuê bao trả trước.




2. Lấy dữ liệu đầu vào

Dữ liệu về địa điểm

Với nhóm các dữ liệu liên quan đến địa điểm và đặc điểm di chuyển. Nhóm dữ liệu này đặc biệt quan trọng, nó thể hiện tính ổn định trong công việc và nơi ở, địa bàn hoạt động hiện tại của thuê bao.

-  Những dữ liệu này thể hiện tốt hơn rất nhiều các dữ liệu định danh về địa điểm:
- Dự đoán nơi ở, nơi làm việc, nơi nghỉ ngơi theo các mốc thời gian đã chia:
 - Địa chỉ thường lui tới nhất trong tuần(khả năng cao là nơi làm việc)
 - Địa chỉ thường lui tới nhất ban tối trong tuần(khả năng cao là nơi ở hoặc là 1 địa điểm yêu thích nào đó)
 - Địa chỉ thường lui tới nhất vào cuối tuần (khả năng là nơi nghỉ ngơi yêu thích)





-  Tính ổn định của thuê bao dựa trên tần suất tại các địa điểm:
- Số lần chuyển từ xã này sang xã khác trong cùng huyện.
 - Số lần chuyển từ huyện này sang huyện khác trong cùng tỉnh
 - Số lần chuyển từ tỉnh này sang tỉnh khác
 - Số các địa điểm di chuyển: Thể hiện bằng số lần đổi cell theo mỗi time window
 - Tổng số các địa điểm hoạt động.
 - Số ngày thuê bao xuất hiện tại địa điểm đó



2. Lấy dữ liệu đầu vào

Dữ liệu về quan hệ xã hội

Để phân tích các mối quan hệ thuê bao, mô hình có đưa vào một số dữ liệu như:

-  Đếm số các thuê bao phân biệt có liên lạc với thuê bao đã cho: Thể hiện các mối quan hệ xã hội của thuê bao đó
-  Đếm số các cuộc gọi phát sinh 2 chiều trong vòng 1h. Thể hiện các mối quan hệ có tính tương tác cao.
-  Đếm số các cuộc gọi phát sinh 2 chiều trong vòng 1h và có tính ổn định cao theo các tuần: Thể hiện mối quan hệ thân thiết của thuê bao đó.
-  Đếm số các thuê bao thường xuyên gọi/nhận tới những thuê bao trong mối quan hệ với thuê bao đã cho: Thể hiện các quan hệ cấp 2 với thuê bao

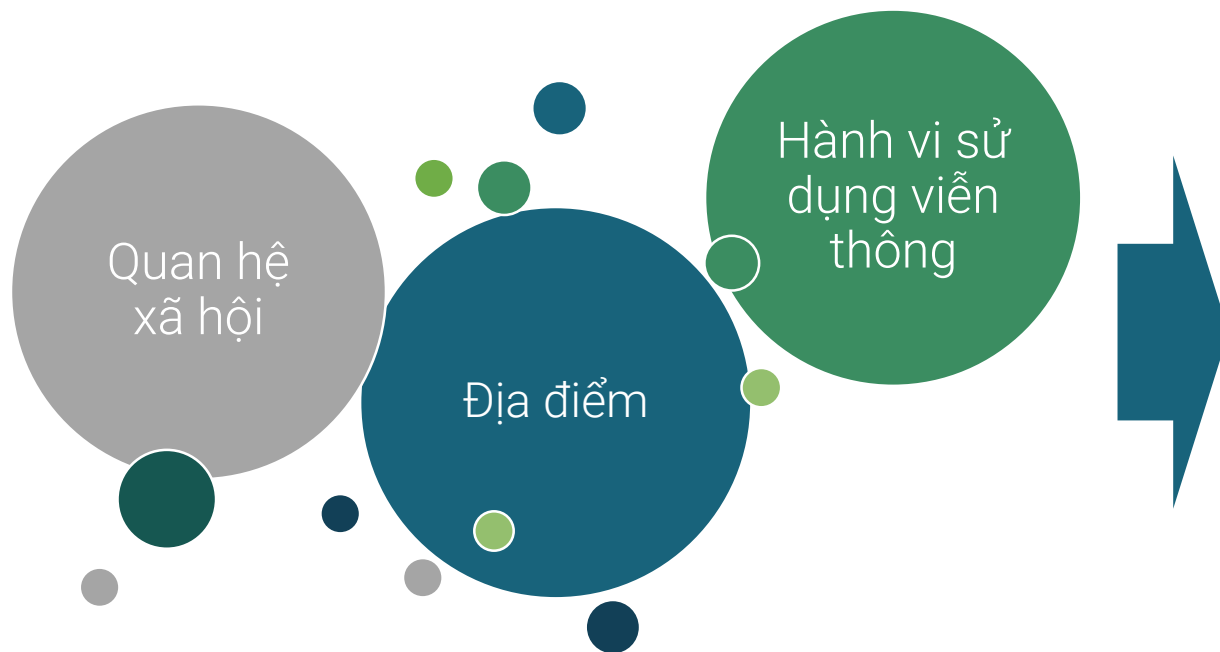
➤ Các dữ liệu này thể hiện mối quan hệ gia đình, xã hội của thuê bao ở nhiều cấp độ khác nhau. Ngoài ra còn giúp mô hình tìm ra những mối liên hệ với cá nhân có nợ xấu



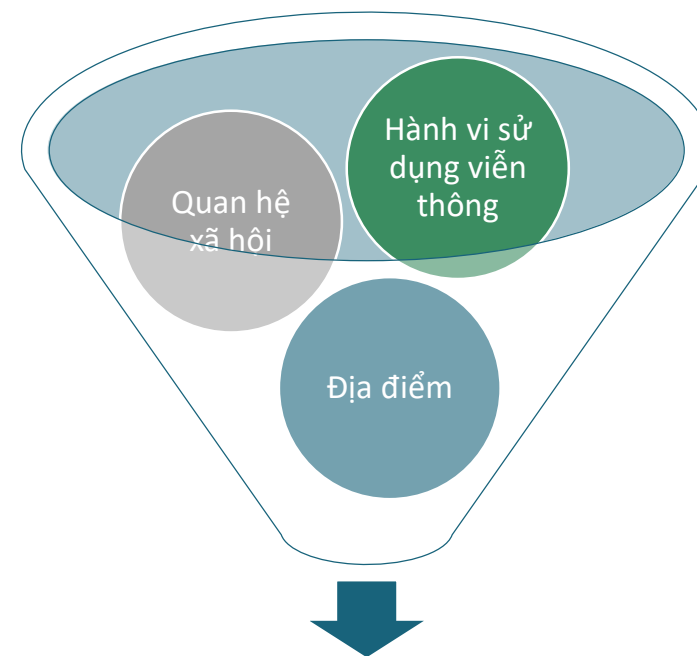
2. Lấy dữ liệu đầu vào

Tư duy lựa chọn dữ liệu

Các dữ liệu sơ cấp ban đầu được tư duy và xây dựng dựa vào các bài báo nghiên cứu uy tín trên thế giới*



Các biến thứ cấp được xây dựng từ dữ liệu sơ cấp theo hướng tư duy tối đa hóa để giúp trích xuất được nhiều thông tin một cách toàn diện về thuê bao, nhằm hỗ trợ cho quá trình học máy.



Dữ liệu được đưa vào mô hình

Các dữ liệu không có giá trị để chấm điểm, sẽ bị loại ra trong quá trình xử lý & làm sạch, hoặc được mô hình tự sàng lọc, lựa chọn.

3

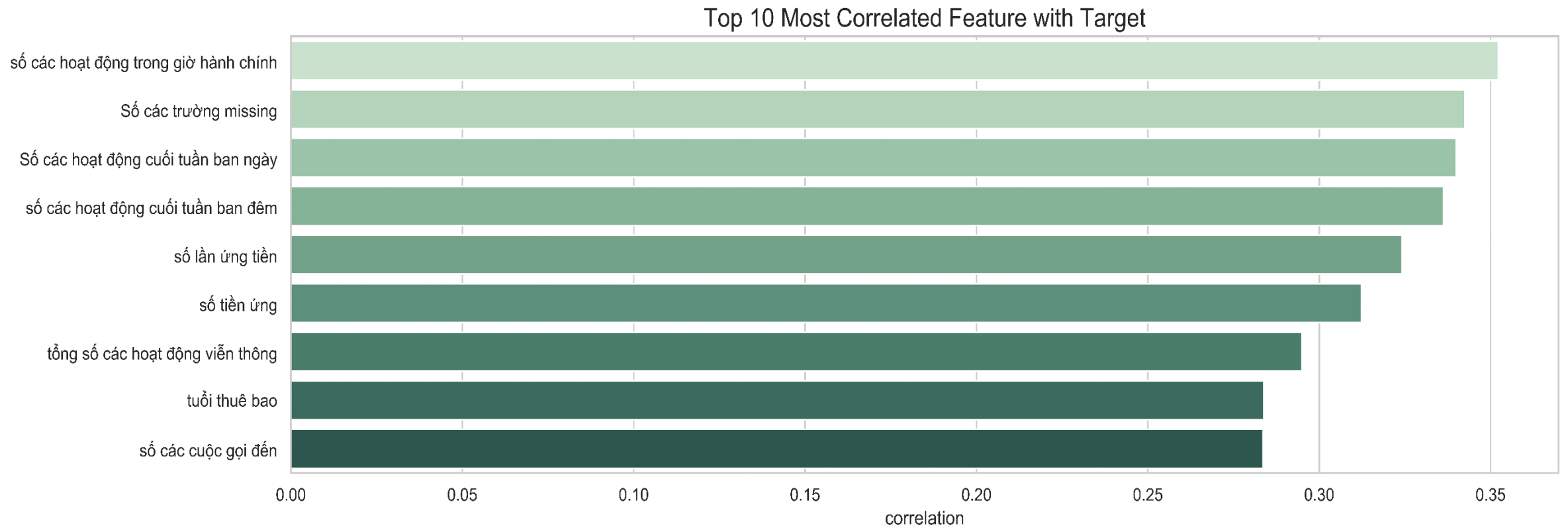
Xử lý & phân tích dữ liệu

Data processing & analysis



3. Xử lý & phân tích dữ liệu

Kiểm tra tương quan



©2020 PTDL - TTCN - VDS

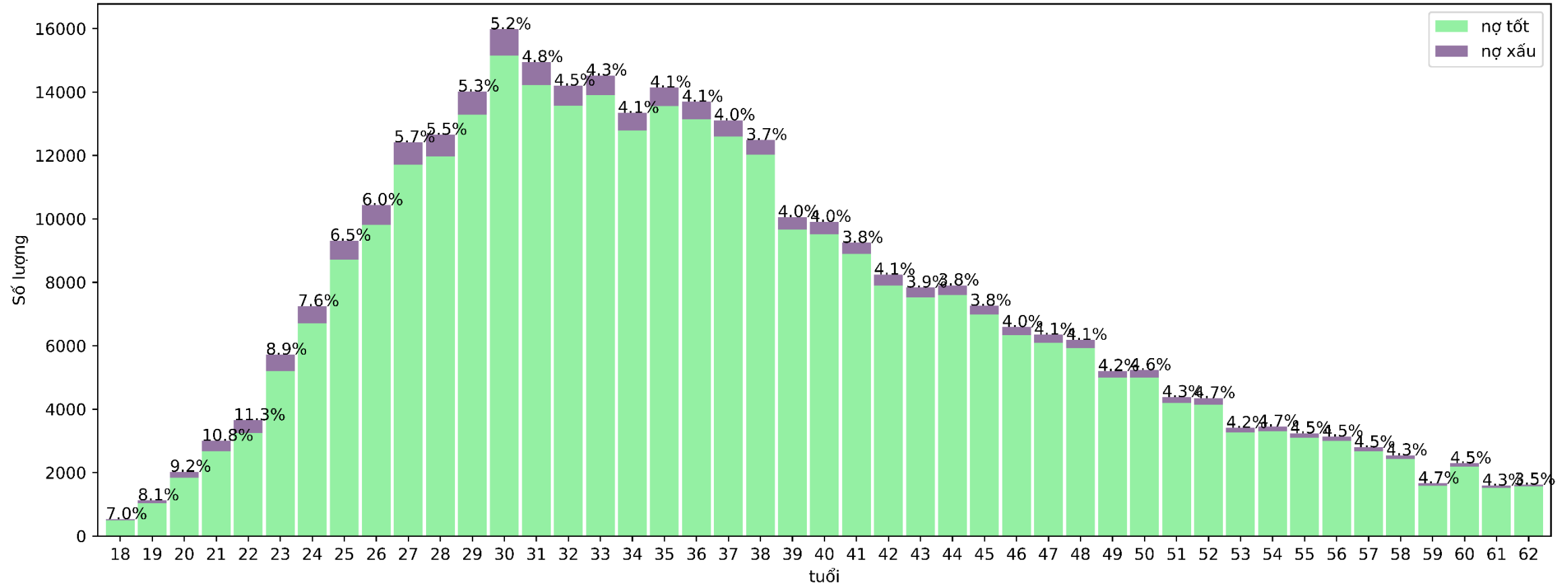
➤ 10 biến hàng đầu tương quan tốt nhất với nhãn nợ xấu/tốt có hệ số giao động từ 0.28 đến 0.35. Có thể thấy các trường này có mối liên hệ tuyến tính khá tốt với các nhãn nợ xấu/tốt.



3. Xử lý & phân tích dữ liệu

Phân bố giá trị biến

Phân Bố Của Nợ Tốt-Xấu Tuổi



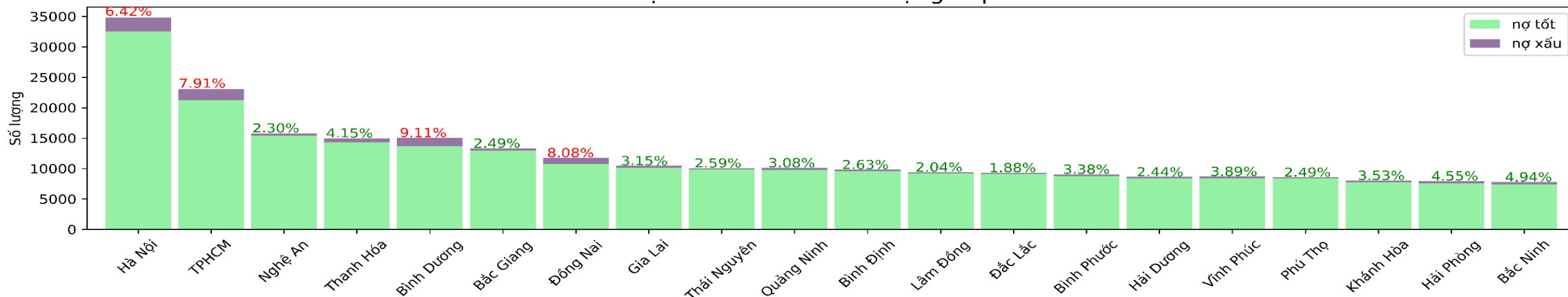
➤ Tỷ lệ nợ xấu cao ở nhóm độ tuổi dưới 32 và thấp hơn nhiều ở nhóm còn lại.



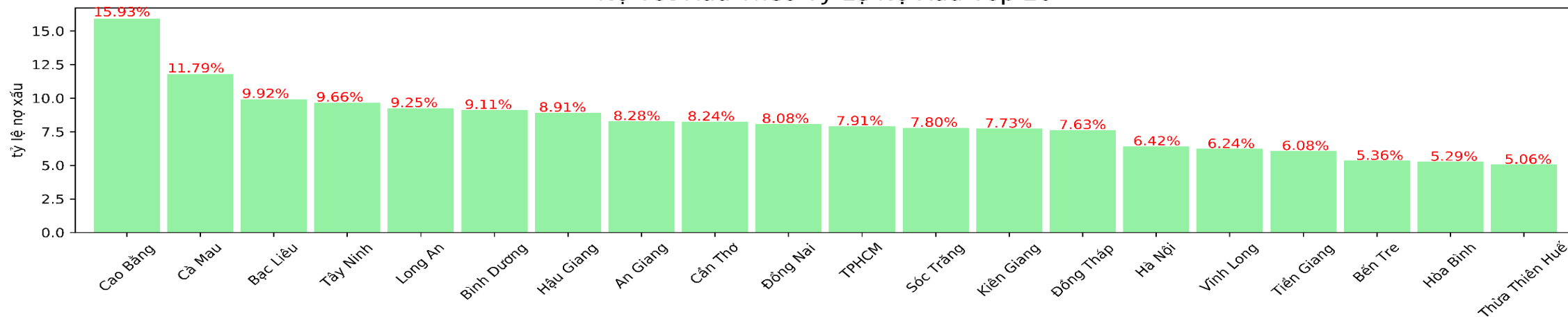
3. Xử lý & phân tích dữ liệu

Phân bố giá trị biến

Nợ Tốt-Xấu Theo Số Lượng Top 20



Nợ Tốt-Xấu Theo Tỷ Lệ Nợ Xấu Top 20



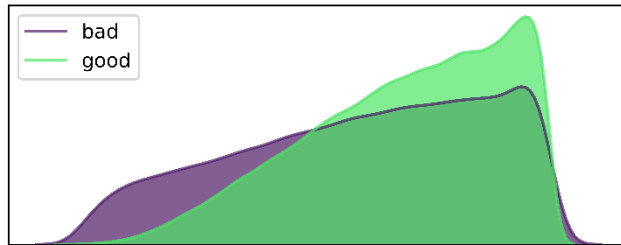
- Các thuê bao phát sinh khoản vay phân bố nhiều ở các tỉnh/thành phố lớn như Hà Nội, TP Hồ Chí Minh, Nghệ An, Thanh Hóa, trong đó Hà Nội, HCM, Bình Dương, Đồng Nai có tỷ lệ nợ xấu cao đột biến.
- Các tỉnh Cao Bằng, Cà Mau, Bạc Liêu,... xếp top nợ xấu tuy nhiên số lượng quan sát thấp, độ tin cậy không cao



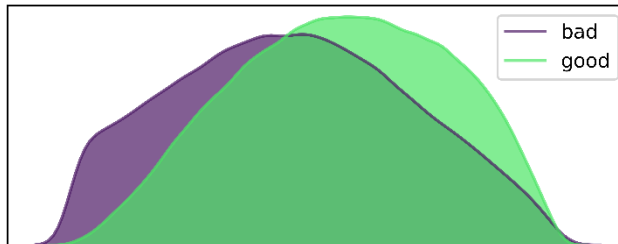
3. Xử lý & phân tích dữ liệu

Phân bố giá trị biến

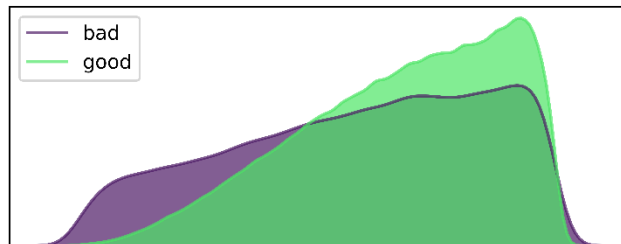
Phân Phối Trung Bình Số Ngày Hành vi sử dụng viễn thông Vào Buổi Tối & Trong Tuần



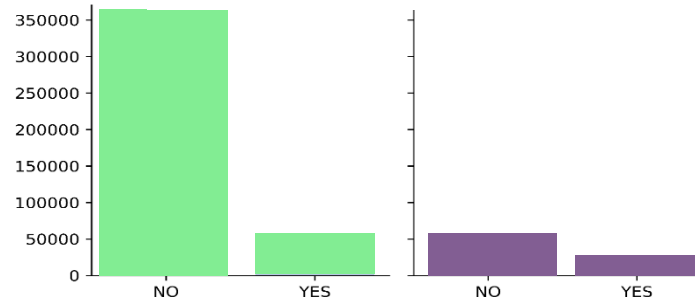
Phân Phối Trung Bình Tần Suất Tới Địa Chỉ Hay Đến Nhất Vào Buổi Tối & Cuối Tuần



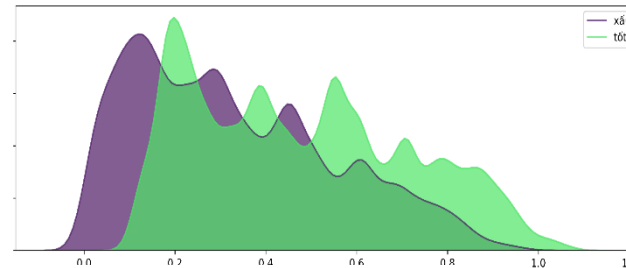
Phân Phối Trung Bình Số Ngày Hành vi sử dụng viễn thông Vào Buổi Tối & Cuối Tuần



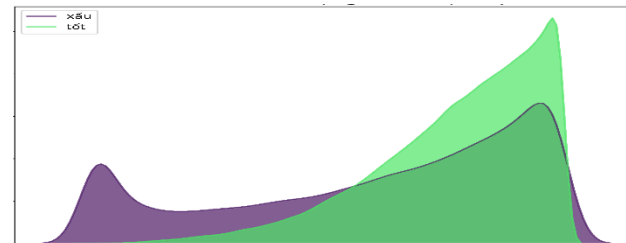
Phân Phối Thuê Bao Có Ít Nhất 01 Lần Thay Đổi Địa Điểm Làm Việc



Phân Phối Tuổi Thuê Bao Viễn Thông (Đã Biến Đổi Chuẩn Hóa)



Phân Phối Số Lượng Cuộc Gọi/Tháng Của Thuê Bao



Phân tích một số biến quan trọng nhận thấy phân phối giá trị có tính phân biệt về tín dụng xấu, tốt



Dữ liệu viễn thông có khả năng để xếp hạng tín dụng các thuê bao

Nợ xấu

Nợ tốt

4

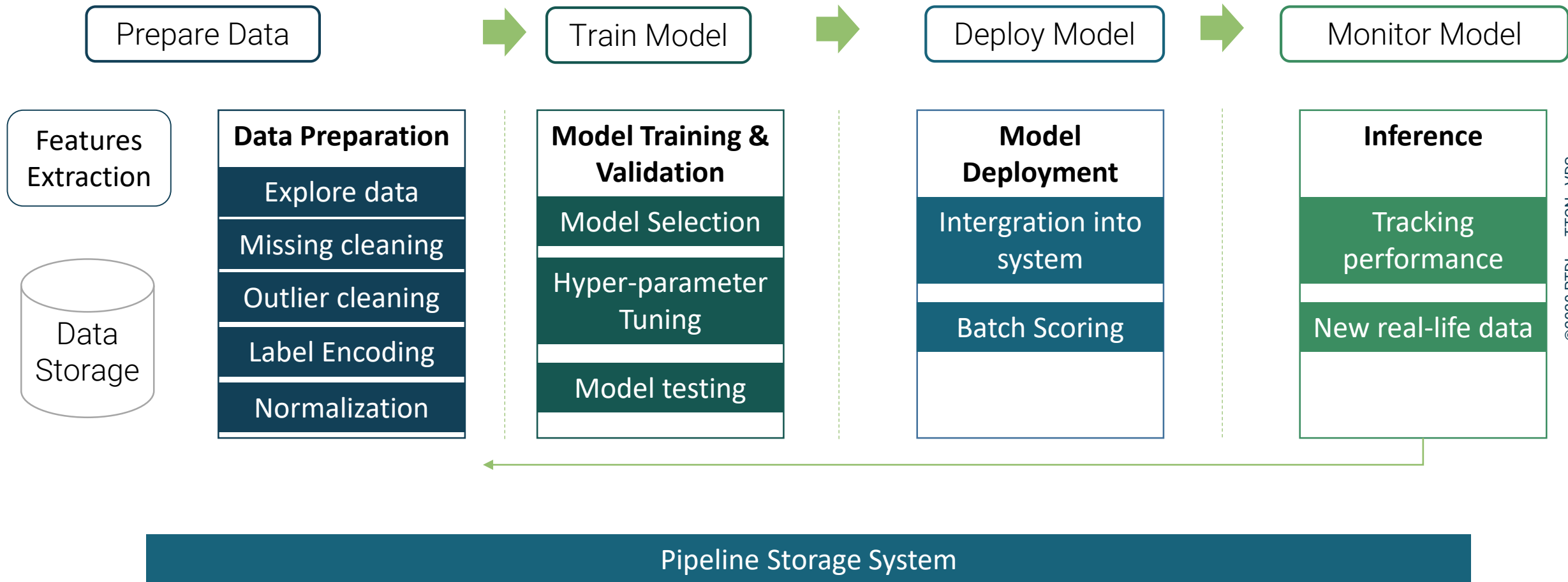
Xây dựng mô hình

Data modelling



4. Xây dựng mô hình

Toàn cảnh luồng mô hình

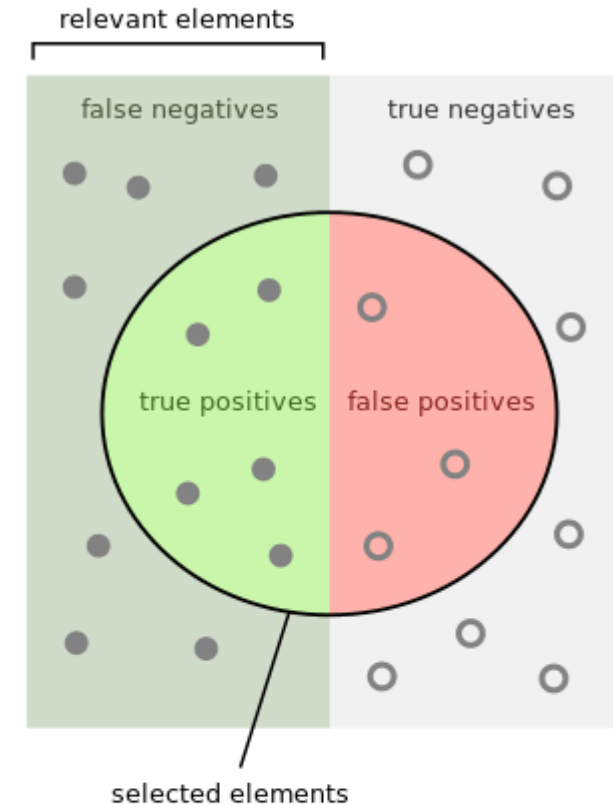




4. Xây dựng mô hình

Tiêu chí đo lường kết quả

- Gini: Giá trị này giao động trong khoảng từ -1 đến 1 tương ứng với việc đoán sai hoàn toàn và đúng tuyệt đối.
- Recall: Tỷ lệ dự báo đúng nhãn xấu trên tổng số nhãn xấu.
- Balanced Accuracy: Trung bình có trọng số của tỷ lệ dự báo đúng các nhãn trên tổng số dữ liệu. Nhãn càng hiếm thì trọng số càng cao. Tiêu chí này hay được dùng trong trường hợp bộ dữ liệu có nhãn rất lệch, khi đó, việc đoán đúng các nhãn hiếm là rất quan trọng.



How many selected items are relevant?

Precision =



Activate Windows

How many relevant items are selected?

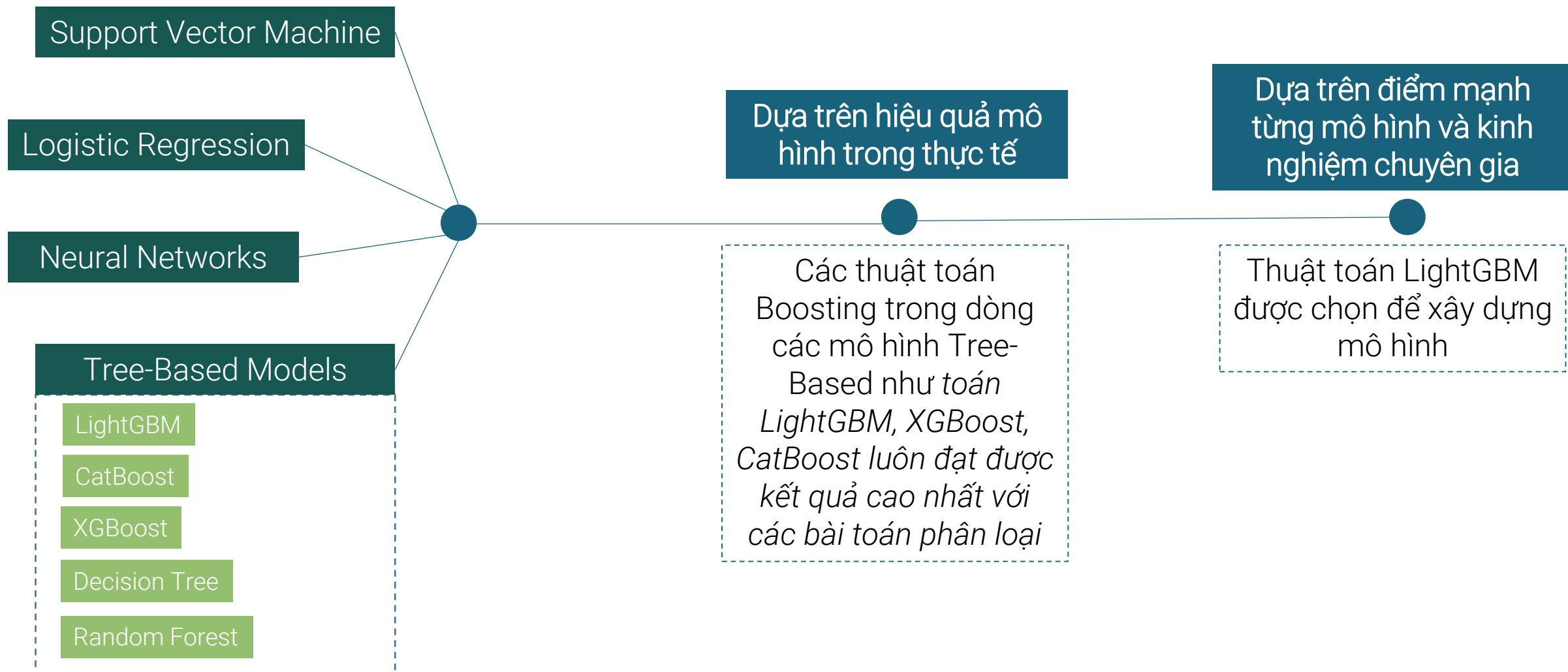
Recall =





4. Xây dựng mô hình

Lựa chọn thuật toán





Ưu điểm thuật toán

Các điểm mạnh của LightGBM

Tốn ít tài nguyên, dễ dàng chạy trên Hadoop Spark (server) hoặc máy cá nhân (local).

Kết quả tốt trên nhiều bộ dữ liệu khác nhau với rất ít cài đặt về tham số.

Hạn chế tốt vấn đề overfitting.

Dễ dàng xử lý dữ liệu bị thiếu, dữ liệu không phải dạng số, và tỷ lệ nhãn lệch.



Các ưu điểm của K-Folds*

Áp dụng K-Folds bên cạnh LightGBM nhằm mục đích đưa ra kết quả mang tính khách quan, ổn định nhất

Hạn chế tối đa các yếu tố nhiễu do quá trình cắt tập train và test gây ra.

Cách thức: Chia ngẫu nhiên tập dữ liệu thành 5 phần bằng nhau. Lần lượt train trên 4 phần và test kết quả trên tập còn lại




4. Xây dựng mô hình


Kết quả dự báo

	Dự báo TB xấu	Dự báo TB tốt	Tổng
TB xấu thực tế	16,483	4,327	20.810
TB tốt thực tế	211,030	443,801	654.831
Tổng	227.513	448.128	675.641

- Gini = 0,72 sai số trung bình 0,01.
- Balanced accuracy = 73,49 % sai số trung bình 0,06%
- Recall score = 79,21% sai số trung bình 0,04%

Có thể thay đổi cấu trúc của ma trận nhầm lẫn (confusion matrix) bằng cách thay đổi ngưỡng mô hình (threshold) về nợ tốt/xấu để phù hợp hơn với từng hoàn cảnh:

 **Hạn chế tối đa nợ xấu:** Trong trường hợp ta không có nhiều vốn, hoặc có khẩu vị rủi ro thấp sẽ đỡ ngưỡng cao nhưng hạn chế là có nhiều nợ tốt không được xét vay do tiêu chuẩn khắt khe.

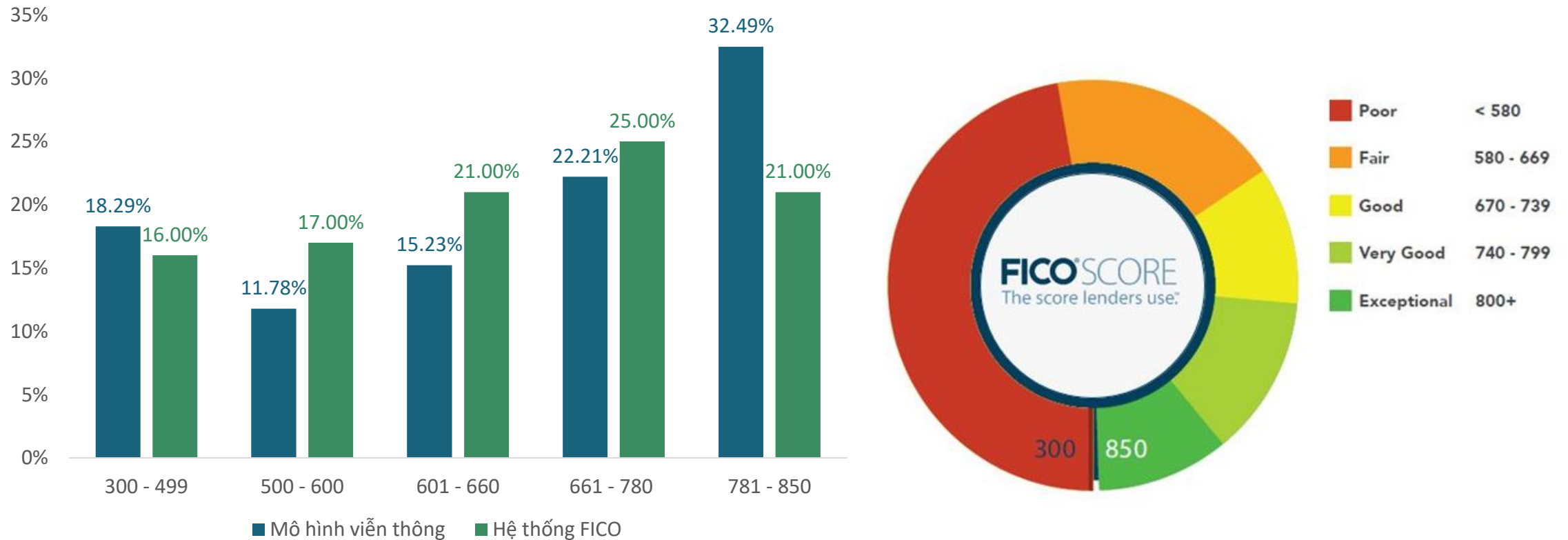
 **Tăng tỷ lệ giải ngân:** Trong trường hợp ta có nguồn vốn dồi dào, và khẩu vị rủi ro cao thì xét ngưỡng thấp nhưng hạn chế là tỷ lệ nợ xấu có nguy cơ tăng cao.



4. Xây dựng mô hình

So sánh phổ điểm mô hình

Mô hình dự báo đưa ra ngưỡng điểm xếp hạng tín dụng khách hàng theo chuẩn FICO (300 – 850) với 05 mức:



➤ Phân phối phổ điểm do mô hình viễn thông dự báo có nhiều điểm tương đồng với phân phối phổ điểm của hệ thống FICO*, cho thấy tính tin cậy của mô hình

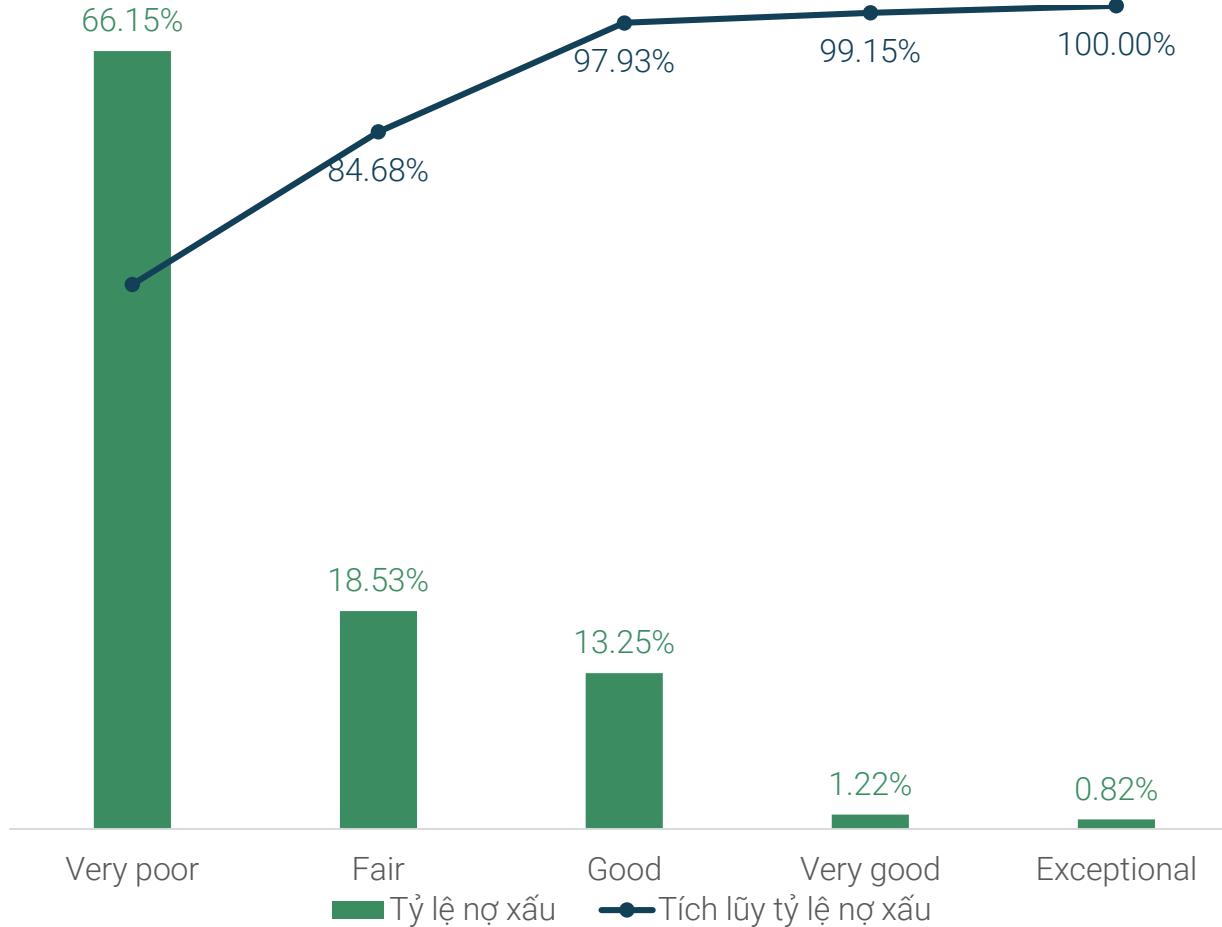
* [Truy cập đường link xem thêm về tính phổ biến của hệ thống FICO](#)



4. Xây dựng mô hình

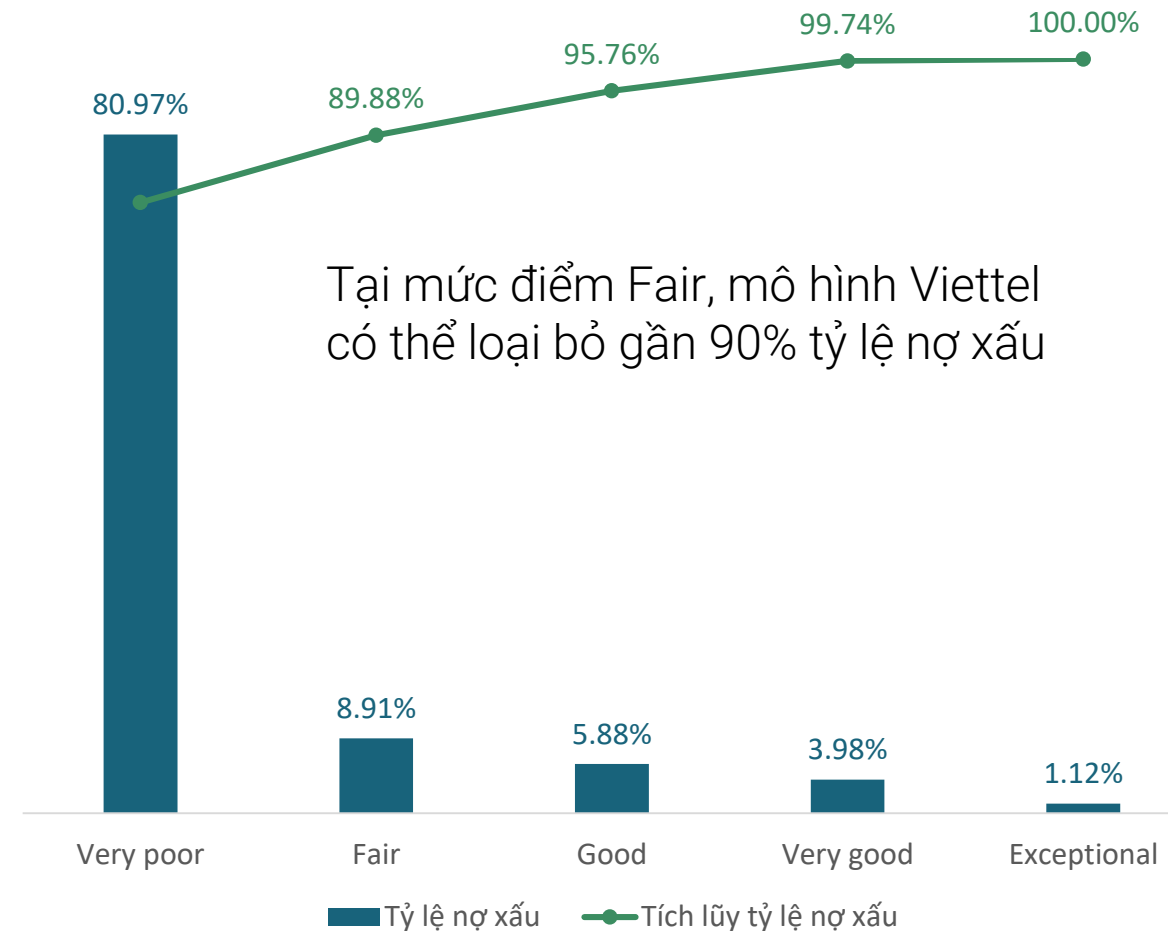
So sánh tỷ lệ nợ xấu

Tỷ lệ nợ xấu theo mô hình
Trusting Social



Tài liệu Trusting Social năm 2016

Tỷ lệ nợ xấu theo mô hình Viettel



Tại mức điểm Fair, mô hình Viettel có thể loại bỏ gần 90% tỷ lệ nợ xấu

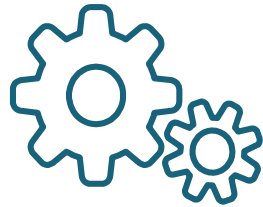


4. Xây dựng mô hình

Đánh giá nguy cơ tiềm ẩn

Mô hình có kết quả khá tốt trên tập dữ liệu (train và test) đang có, mang tính khả thi cao. Tuy nhiên, vẫn tiềm ẩn những nguy cơ sau:

Rò rỉ dữ liệu
Data leakage*



Nhãn nợ xấu được thu thập sau khi việc nợ xấu đã xảy ra do đó hành vi có thể khác với việc dự báo nợ xấu khi việc này chưa xảy ra.

Dữ liệu không đại diện



Tập dữ liệu đang sử dụng để xây mô hình có thể không đại diện cho phân bố chung của tập lớn (ở đây là tất cả thuê bao có nhu cầu vay vốn trên thị trường).

* Việc này có thể kiểm tra được bằng dữ liệu, tuy nhiên hiện tại dữ liệu chưa đủ dài để có thể nhìn về quá khứ của các thuê bao này và phát hiện những sai khác về thói quen viễn thông nếu có.

5

Phương án triển khai

Pilot model



5. Phương án triển khai

Các phương án đề xuất

Ứng dụng trực tiếp mô hình chấm điểm vào tập khách hàng trên Easyvay hiện tại của ViettelPay

Đề xuất 01: Thay thế dần dần điểm của Trusting Social




- + Tùy thuộc vào tình hình kinh doanh, ta có thể **thay thế dần dần từ 10%, 20% đến 50%** tập khách hàng Easyvay bằng mô hình của Viettel
- + Lưu lại dữ liệu vay nợ của 03 tập - tập A: Chấm bằng mô hình của Viettel. Tập B: Chấm bằng mô hình đối tác. Tập C: Rút ngẫu nhiên một ít mẫu cho vay mà không chấm điểm.
- + So sánh hiệu quả trên các tập A, B, C.

Đề xuất 02: Chạy song song hai luồng của Trusting Social và Viettel



- Chạy song song 2 luồng mô hình:
- + Luồng của Trusting Social để cho ra kết quả phục vụ mục đích kinh doanh.
 - + Luồng ViettelPay để có kết quả so sánh với mô hình Trusting Social. Sẽ triển khai bằng luồng của ViettelPay khi hiệu quả đạt đến ngưỡng cho phép

 Mặc dù mô hình đã hoạt động ổn định, có độ chính xác tương đối tốt khi chạy trên dữ liệu hiện tại, nhưng khi đưa vào thực tế hoạt động cần có các biện pháp và chỉ số đo đạc cụ thể để đánh giá hiệu quả mô hình.



PHỤ LỤC BÁO CÁO

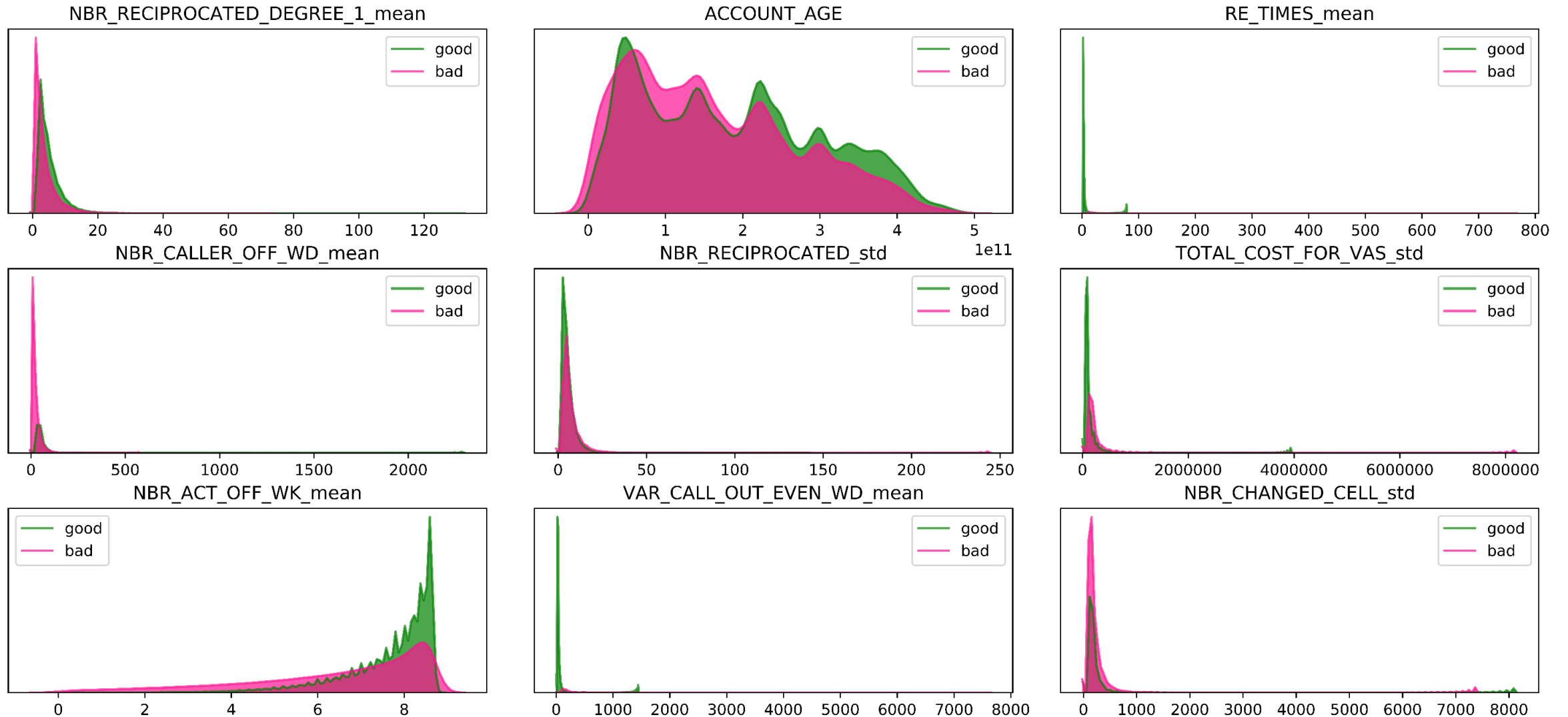
Các nghiên cứu tham khảo

- Lane, M., Carpenter, L., Whitted, T., Blinn, J.: Scan line methods for displaying parametrically defined surfaces. *Communications ACM* 23(1) (1980)
- Ding, C.H.Q., Peng, H.: Minimum redundancy dữ liệu selection from microarray gene expression data. *J. Bioinformatics and Comp. Biol.* 3(2) (2005) 185–206
- Prediction of Socioeconomic Levels Using Cell Phone Records
- Credit Scoring for Good Enhancing Financial Inclusion with Smart(2019)
- Business Intelligence Applications and Data Mining Methods in Telecommunications: A Literature Review https://core.ac.uk/tải_xuống/pdf/71542208.pdf
- And many other at: <http://confluence.digital.vn/x/ZyRDAg>

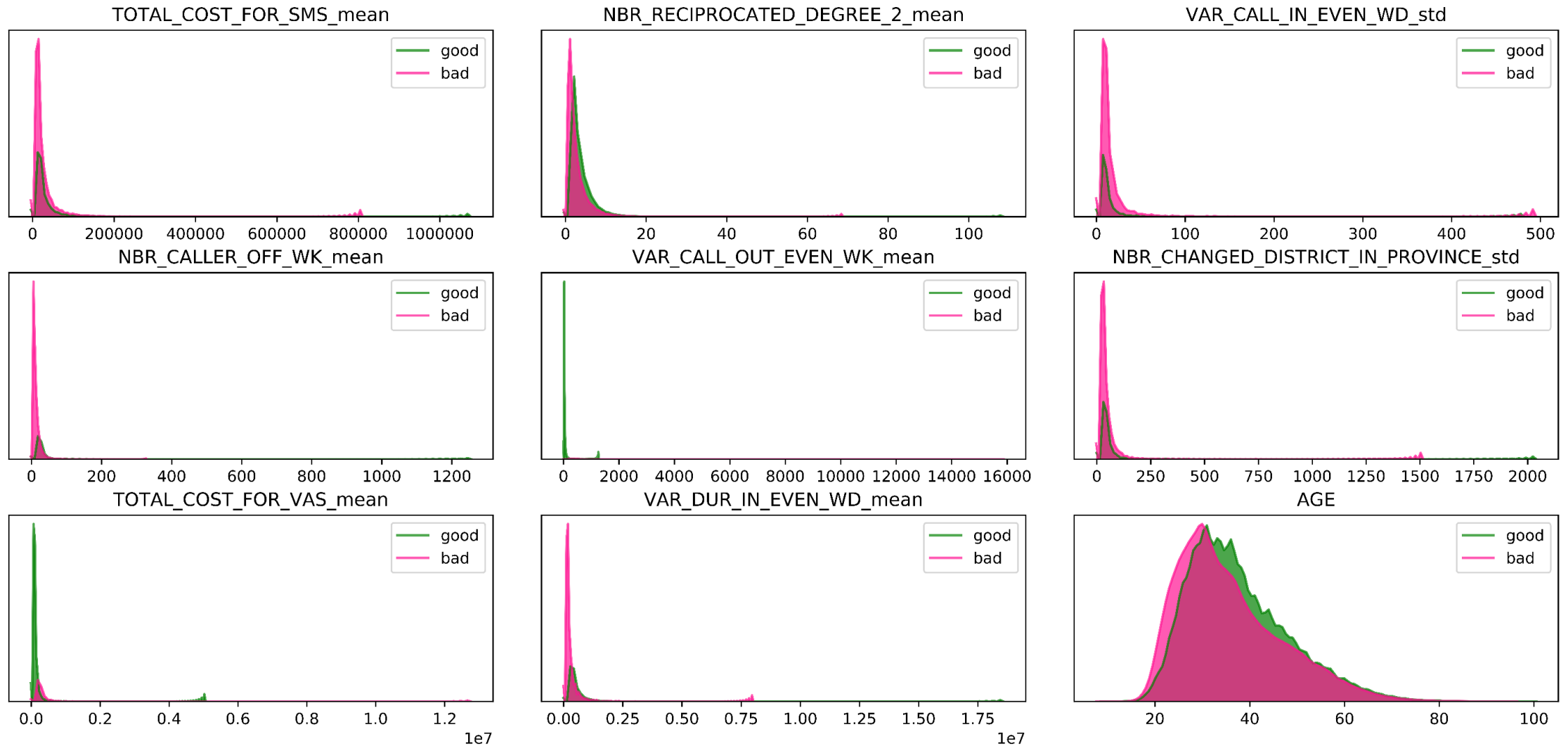
Một số chú thích

- Việc sử dụng dữ liệu cho 5 tháng thay vì 3 tháng có thể làm tăng khoảng 0.03 gini index. Tác giả đề xuất sử dụng dữ liệu 6 tháng liên tiếp để tăng sự ổn định.
- Tiêu chuẩn dropout khỏi mô hình: Các thuê bao có ít nhất 1 tháng không phát sinh giao dịch viễn thông. Khi chạy thực tế, có thể xuất điểm tối thiểu với các thuê bao phát sinh giao dịch viễn thông gián đoạn(phát sinh tháng t nhưng không phát sinh tháng $t+1$)
- Với các **numeric dữ liệu**: Nhóm thành từng nhóm theo các tháng khác nhau, biến thành các dữ liệu là mean và std của các dữ liệu này. Các làm này đã làm tăng gini index khoảng 0.015 so với việc dùng các dữ liệu gốc.
- Với các **categorical dữ liệu**: Encoder bằng label encoder thông thường. Theo kinh nghiệm của tác giả, việc encoder bằng các phương pháp khác như onehotencoder, target encoder không cải thiện được accuracy nếu dùng thuật toán boosting.

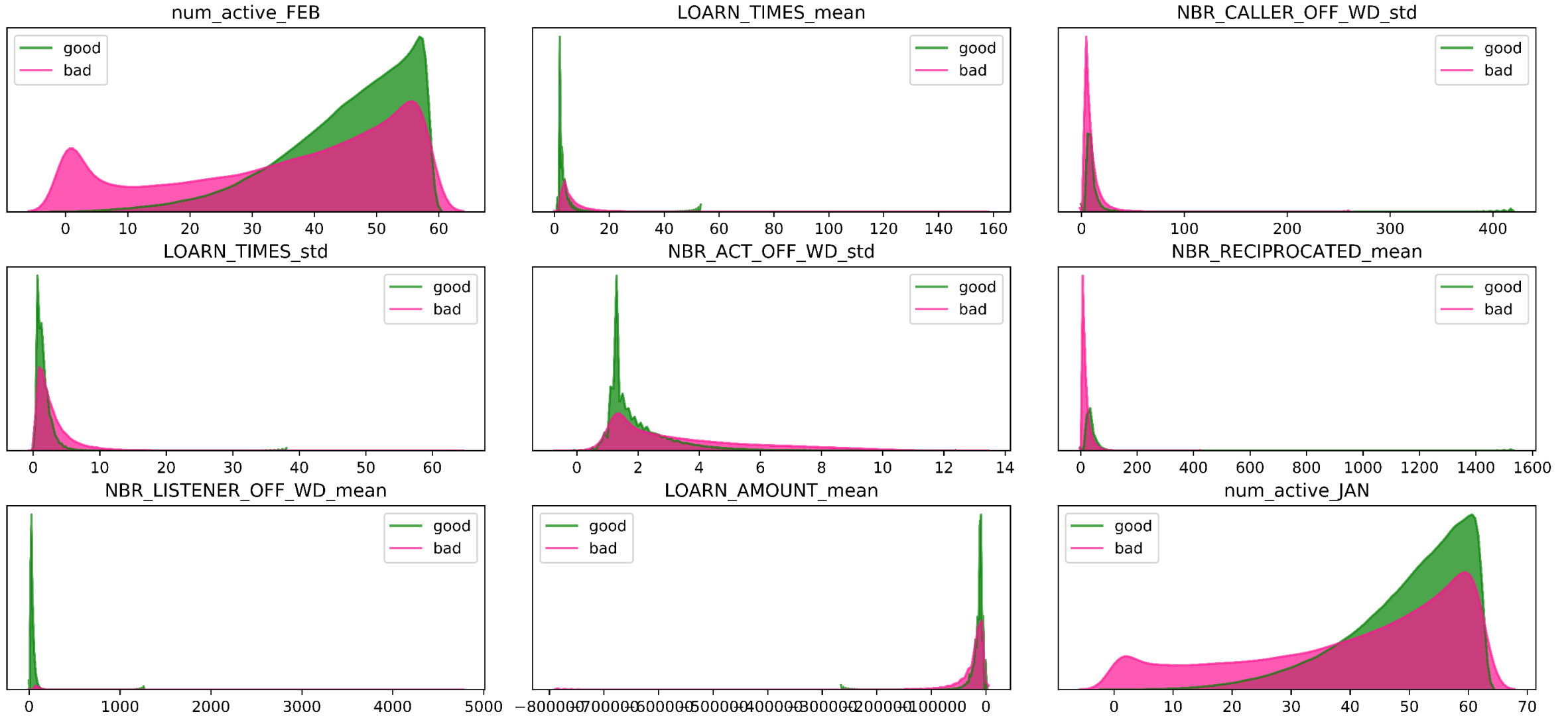
Phân bố giá trị biến



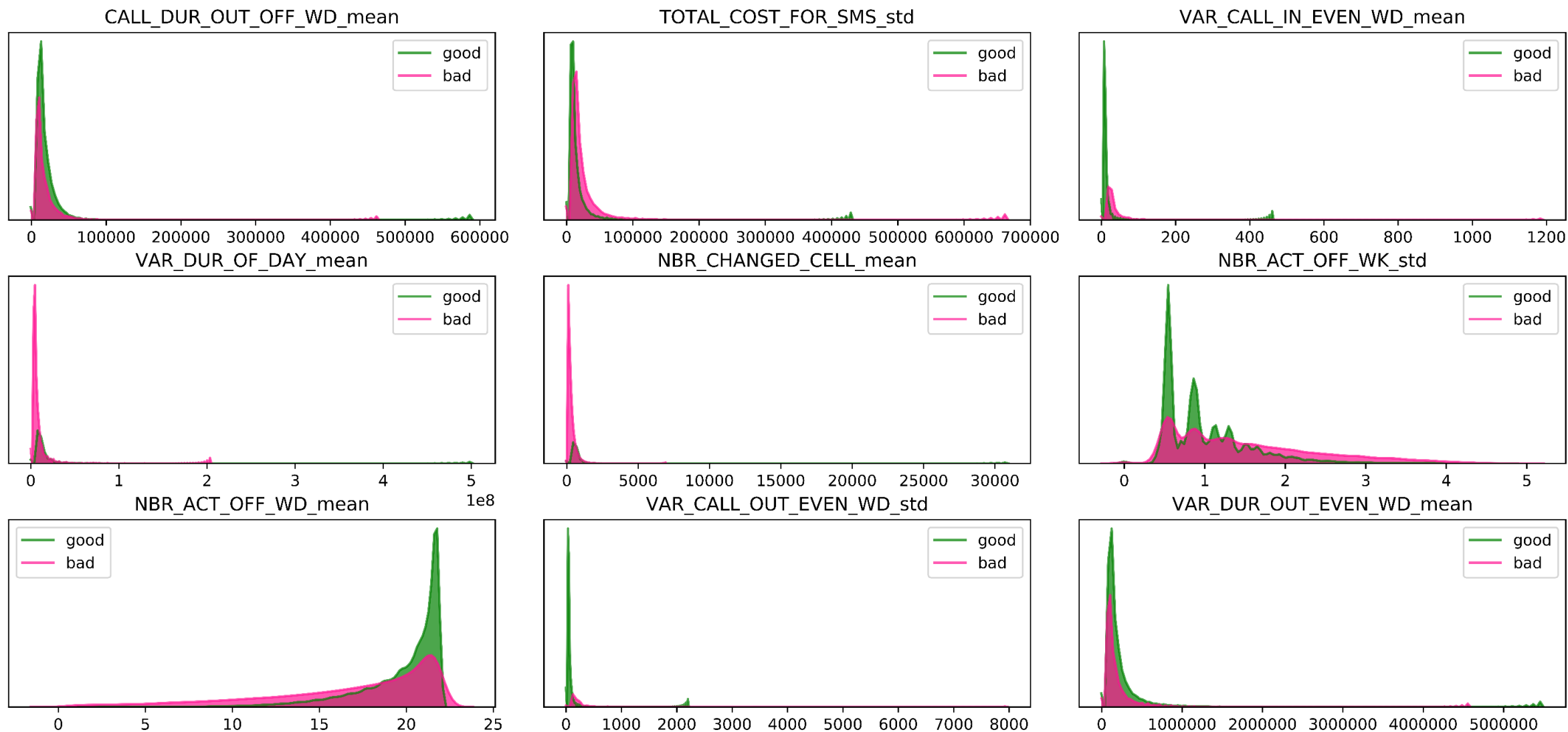
Phân bố giá trị biến



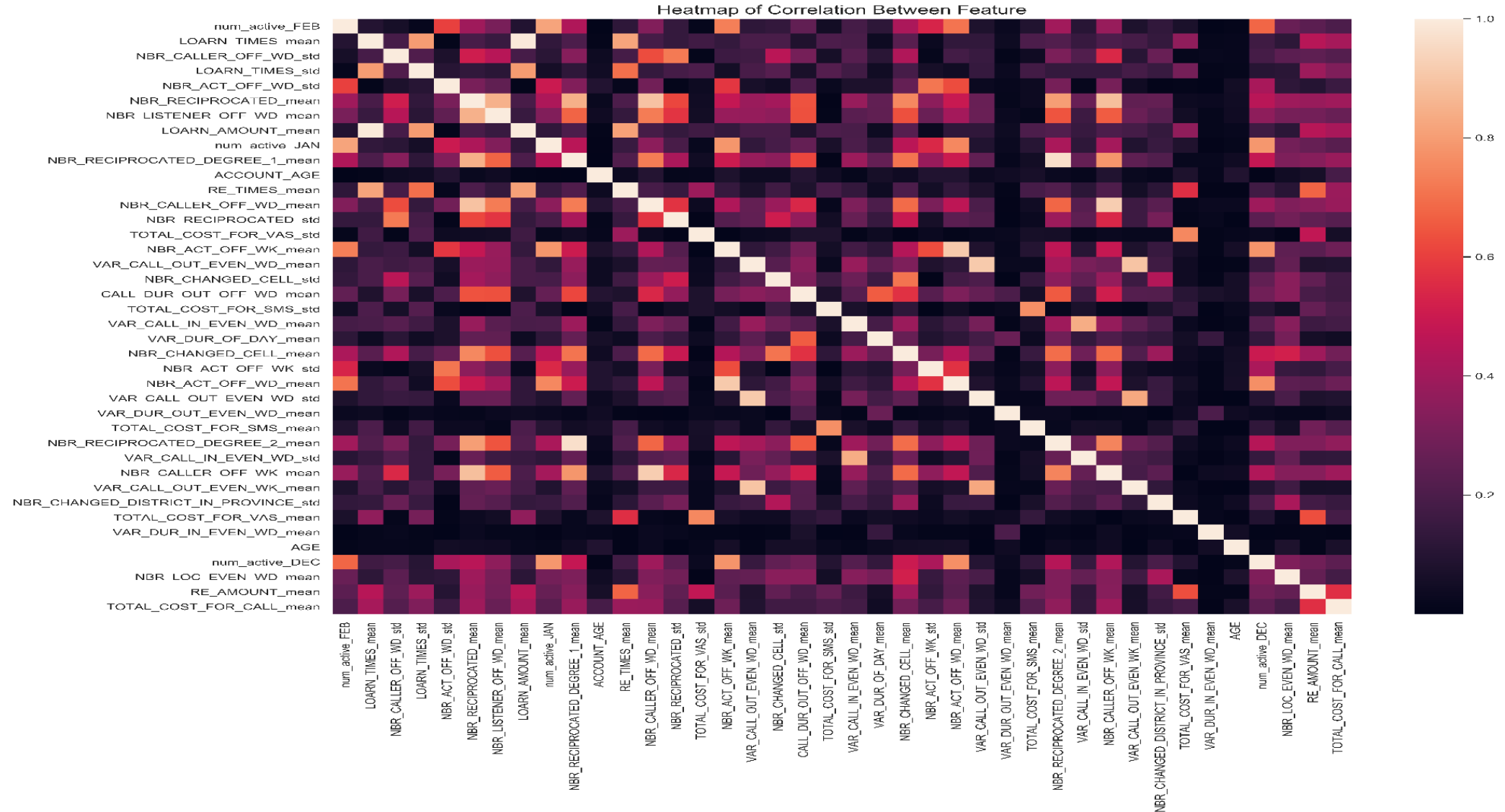
Phân bố giá trị biến



Phân bố giá trị biến



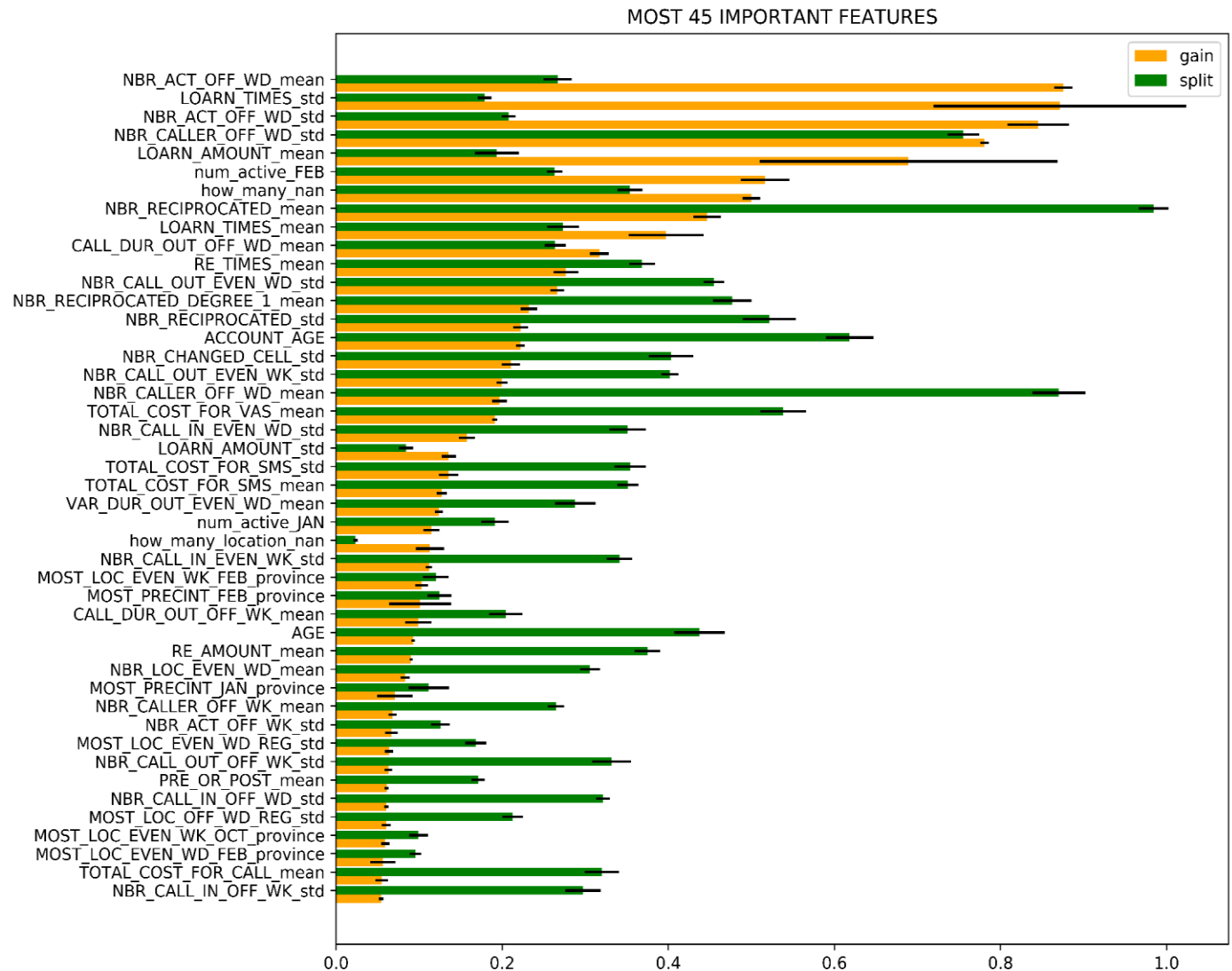
Bản đồ nhiệt hệ số tương quan



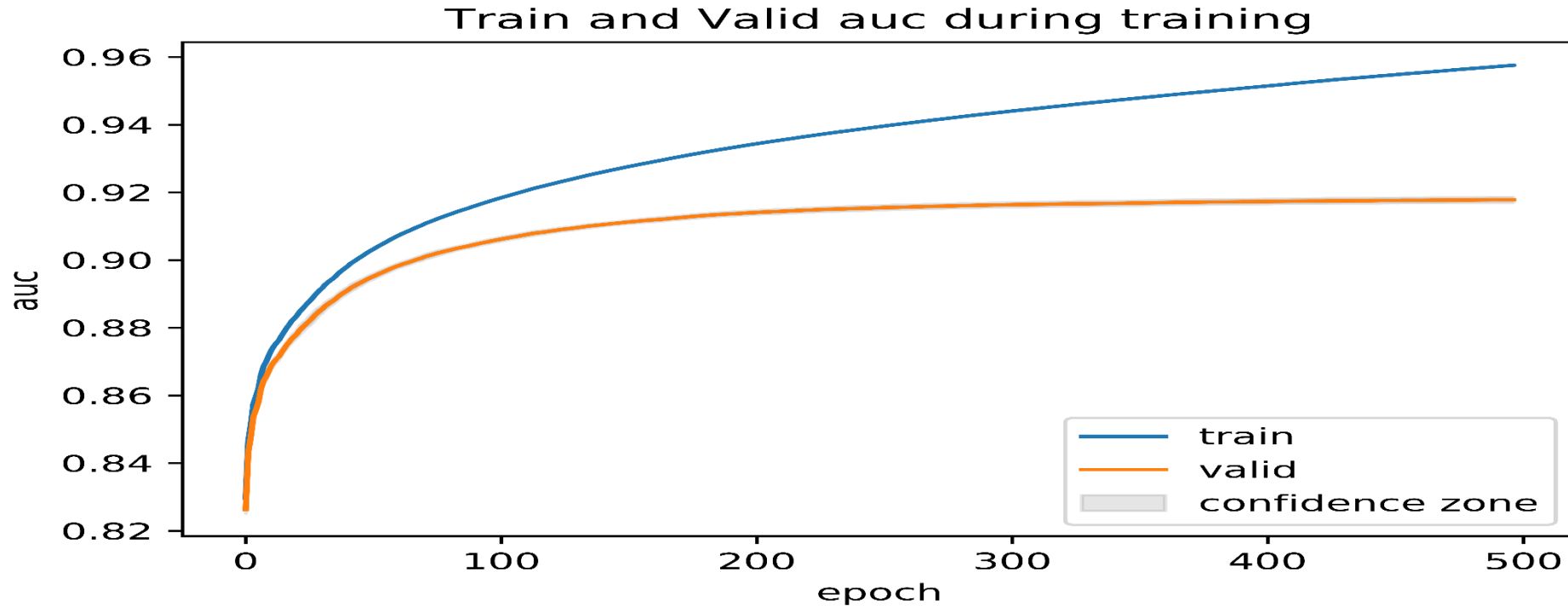
Đánh giá kết quả

Tính hợp lý của mô hình thể hiện một phần ở Top những danh sách quan trọng nhất. Ở đây chúng ta có:

- Nhóm dữ liệu liên quan đến tần suất gọi trong giờ hành chính.
- Nhóm dữ liệu liên quan đến tần suất ứng tiền viễn thông.
- Nhóm liên quan đến các tương tác 2 chiều của thuê bao.

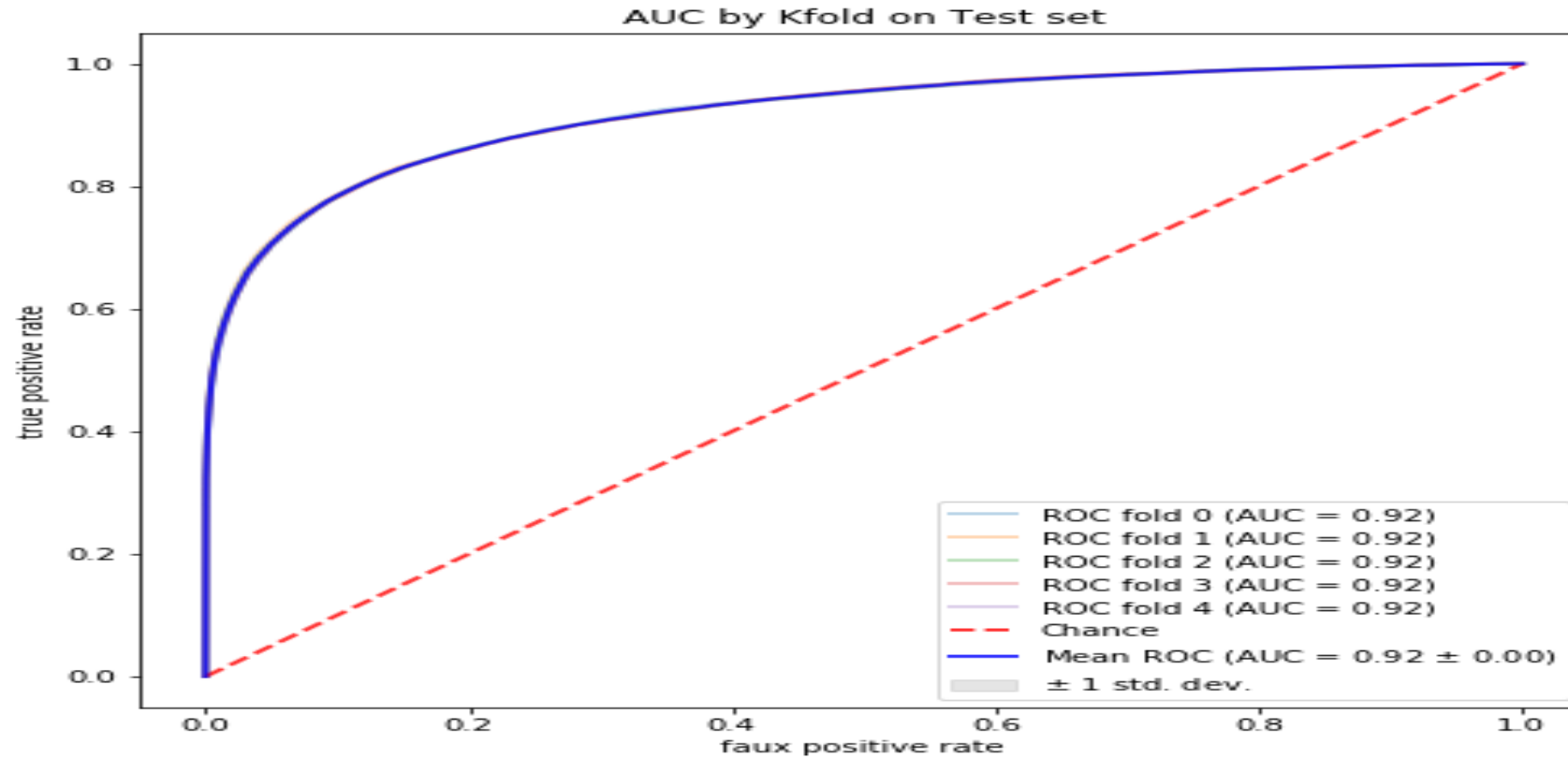


Đánh giá kết quả



Các mô hình thực hiện trên các tập train và test khác nhau đều cho ra kết quả có độ lệch chuẩn rất nhỏ, chứng tỏ tính ổn định rất cao của các mô hình này, do dataset đã đủ lớn

Đánh giá kết quả



Kết quả của mô hình là rất tốt trên các tập test khác nhau, thể hiện ở độ ổn định cao và sai số trung bình thấp

Tính chưa ổn định của kết quả

Khi làm việc với các dataset với các tỷ lệ nhãn xấu/tốt khác nhau có thể cho ra các kết quả mô hình khác nhau.

Trong bảng dưới đây, khi nhãn xấu giảm xuống, gini và các metric khác cũng giảm

	Tỷ lệ nhãn xấu = 8.90%			Tỷ lệ nhãn xấu = 4.66%		
	Gini	Recall	Balanced acc	Gini	Recall	Balanced acc
Test 5 folds	0.829+/-0.001	78.6%+/-0.2%	83.1+/-0.1%	0.812+/-0.003	74.83+/-0.2%	81.9+/-0.1%
Extra test	0.825+/-0.001	78.1%+/-0.2%	82.8+/-0.1%	0.807+/-0.002	74.31+/-0.2%	81.7+/-0.09%

**cần kiểm tra kết quả của mô hình trên các loại phân bố khác nhau này để lấy ra mô hình tốt nhất có thể.
Ngoài ra, cũng có thể phải inference trước tỷ lệ nhãn để có phương pháp đánh trọng số nhãn hợp lý.*

Chuyển đổi xác suất vỡ nợ sang phổ điểm FICO

