**One-way locked SIM card.**

# Contents

# Phần: Xây dựng bài toán.

## Phần 1: Bài toán

Dự đoán việc mở khóa thuê bao cho nhóm khách hàng tiềm năng bị khóa SIM.

Vì một vài lí do nào đó (chủ quan & khách quan), mà khách hàng bị khóa thuê bao 1 chiều, nhưng trong số đó lại có khách hàng tiềm năng sinh lợi. Do đó, ta quyết định nên hay không mở (phân loại 2 nhãn). Những khách hàng tiềm năng được mở khóa sẽ có giá trị kinh doanh hơn để đảm bảo không rời bỏ dịch vụ (như sự hỗ trợ mở khóa, giải đáp, cung cấp dịch vụ phù hợp).

## Phần 2: Mô tả Dataset

| Thông tin | Ý nghĩa | Kiểu dữ liệu |
|---|---|---|
| sdt_mahoa | Thuê bao được mã hóa | Nominal |
| ngay_khoa | Ngày thuê bao bị khóa | Datetime |
| ngay_mo_date | Ngày thuê bao được mở | Datetime |
| ngay_mo_khoa | Khoảng thời gian bị khóa | Discrete |
| ngay_nap_tien_gan_nhat | Khoảng thời gian kể từ lúc bị khóa cho đến khi nạp tiền (từ lúc bị khóa đến lúc nạp tiền, do đó, có thể tính đến việc nạp nhiều lần) | Discrete |
| arpu_1 | Trung bình lợi nhuận của từng thuê bao tháng thứ 1 | Continuous |
| arpu_2 | Trung bình lợi nhuận của từng thuê bao tháng thứ 2 | Continuous |
| arpu_3 | Trung bình lợi nhuận của từng thuê bao tháng thứ 3 | Continuous |
| arpu_4 | Trung bình lợi nhuận của từng thuê bao tháng thứ 4 | Continuous |
| arpu_5 | Trung bình lợi nhuận của từng thuê bao tháng thứ 5 | Continuous |
| data_1 | Lượng data sử dụng tháng thứ 1 | Continuous |
| data_2 | Lượng data sử dụng tháng thứ 2 | Continuous |
| data_3 | Lượng data sử dụng tháng thứ 3 | Continuous |
| data_4 | Lượng data sử dụng tháng thứ 4 | Continuous |
| data_5 | Lượng data sử dụng tháng thứ 5 | Continuous |
| call_1 | Số lượng cuộc gọi trong tháng thứ 1 | Discrete |
| call_2 | Số lượng cuộc gọi trong tháng thứ 2 | Discrete |

| | | |
|---|---|---|
| call_3 | Số lượng cuộc gọi trong tháng thứ 3 | Discrete |
| call_4 | Số lượng cuộc gọi trong tháng thứ 4 | Discrete |
| call_5 | Số lượng cuộc gọi trong tháng thứ 5 | Discrete |
| ngay_goi_tong_dai_gan_nhat | Ngày thuê bao gọi tổng đài gần nhất tính từ lúc bắt đầu gọi | Discrete |
| da_tung_dang_ky_goi | Thuê bao đã từng đăng ký gói hay chưa | Nominal |
| ngay_dang_ky_goi_gan_nhat | Ngày thuê bao đăng ký gói gần nhất (kể từ lúc đăng kí gói trước đó đến lúc đăng ký gói hiện tại) | Discrete |
| danh_sach_goi_dang_ky | Các gói mà thuê bao đã đăng ký | Nominal |
| | | |

## Phần 3: Phạm vi

- Thời gian: 1/9/2021 đến 14/10/2021.
- Mẫu dùng cho huấn luyện: Khách hàng được mở khóa và bị khóa SIM trong thời gian này.

## Phần 4: Nhãn Được lựa chọn dựa trên "ngay_mo_date"

- 1: Mở, nếu "ngay_mo_date" có ghi nhận giá trị ngày mở.
- 0: Không mở, nếu "ngay_mo_date" có missing value.
- Có tổng cộng 207789/336887 thuê bao được mở khóa.
- Những thuê bao chưa được mở khóa có thể họ chưa chủ động xin mở, hoặc nằm trong các trường hợp vi phạm.

## Phần 5: Phương án đánh giá

- Lượng người được mở khóa thuê bao phán đoán được phải gần/đủ với lượng thuê bao được mở khóa thực tế (true positive). **Dùng AUC ROC để cực đại hóa true positive.**
- Việc xác định ra nhiều trường hợp false positive (tức không mở nhưng lại dự đoán là mở) cũng cần được quan tâm vì thực tế các thuê bao đó không sinh lời nhưng lại làm ảnh hưởng đến các thuê bao khác, làm ảnh hưởng uy tín dịch vụ. Dùng precision để cực tiểu false positive.
- Phán đoán đúng việc không mở khóa cho các thuê bao đang khóa không quan trọng vì không tác động được gì thêm.
- Phán đoán ra các trường hợp false negative (thực tế cần mở nhưng lại dự đoán không mở ) cần quan tâm, vì có thể làm khách hàng bất mãn, giảm uy tin dịch vụ. Dùng recall để cực tiểu false negative.
- **Tóm lại, sử dụng độ đo AUC ROC và f1 score (để cân bằng recall và precision).**

- **True Positive: Thuê bao cần mở khóa được dự đoán mở đúng.**
- False Positive: Thuê bao không mở khóa nhưng dự đoán mở.
- True Negative: Thuê bao không mở khóa và dự đoán không mở khóa đúng.
- False Negative: Thuê bao cần mở khóa nhưng dự đoán không mở.

| Độ đo | Ý nghĩa | Lợi ích | Hạn chế |
|---|---|---|---|
| Accuracy | Kiểm tra tỉ lệ phán đoán đúng các trường hợp Negative và Positive trên tổng các mẫu dự đoán | Cho thấy khả năng mô hình khi phán đoán với những trường hợp Positive và Negative là đúng và dữ liệu cần cân bằng nhãn | Dữ liệu nhiều ở nhãn Positive nên độ đo không hiệu quả. Mặt khắc, ta không quan tâm True Negative vì phán đoán đúng không phục vụ mục tiêu bài toán |
| Recall | Kiểm tra tỉ lệ đoán đúng các trường hợp Positive trên tổng các Positive thực tế. | Cực tiểu hóa false negative. Tránh tình trạng những khách hàng tiềm năng bị khóa dịch vụ gây ra sự rời bỏ => giảm doanh thu. Chi phí cho các trường hợp false neagtive là cao | Trong trường hợp chi phí mất mát cho false negative không cao (như khách hàng gắn bó dịch vụ vì một lí do cấp thiết nào đó, ta không cần phải làm nhiều chiến dịch thuyết phục) thì việc cực tiểu hóa không quan trọng |
| Precision | Kiểm tra tỉ lệ đoán đúng các trường hợp Positive trên tổng các phán đoán Positive của mô hình | Cực tiểu hóa false positive. Tránh tình trạng các thuê bao Negative đó không sinh lời nhưng lại làm ảnh hưởng đến các thuê bao khác, làm ảnh hưởng uy tín dịch vụ nói chung. Ngoài ra, việc support cho các trường hợp này không lớn. Chi phí bỏ ra để khắc phục cho false positive là cao | Trong trường hợp chi phí do false positive là thấp (dễ có cách thức hạn chế việc làm phiền, spam) |
| F1 | Cân bằng Precision và Recall | | |
| AUC ROC | Kiểm tra khả năng phán đoán của mô hình ở các ngưỡng khác nhau | Càng cao thì việc mô hình phán đoán được Recall cao và FPR thấp. Như vậy, AUC ROC giúp ta tối ưu cả True Positive, False | |

| | | | |
|---|---|---|---|
| | | Negative và false Postive | |

# Phần: Tiền xử lý.

## Phần 1: Kiểm tra duplicate để tránh nhiễu

Ta thấy có 6172 mẫu thuê bao bị trùng lặp, song khi quan sát nhận thấy rằng thời gian bị khóa và mở khóa là khác nhau (những giá trị trong các cột còn lại là như nhau). Ta lý giải là do thuê bao này gặp sự cố khi nhập sai mã thẻ nhiều lần và do đó ta sẽ loại bỏ những duplicate này vì nó gây ra nhiễu khi phán đoán với cùng một mẫu (thuê bao). Vậy thực tế ta có 333642 thuê bao.

## Phần 2: Loại bỏ các thuê bao bị quá nhiều missing value ở các cột numerical

Phương án drop theo cột không khả thi vì missing rate trên các cột sau rất nhiều, do đó có thể làm mất mát thông tin. Ngoài ra, đây là những cột quan trọng thể hiện tiềm năng lợi nhuận của các thuê bao cho mô hình phán đoán, nên ta cần giữ.

| | |
|---|---|
| Lượng missing values trên các cột:<br>apru_1 71370<br>apru_2 8786<br>apru_3 62952<br>apru_4 74899<br>apru_5 79302<br>data_1 71370<br>data_2 8786<br>data_3 62952<br>data_4 74899<br>data_5 79302<br>call_1 71370<br>call_2 8786<br>call_3 62952<br>call_4 74899<br>call_5 79302 | Như hình bên, ta thấy có 71370 thuê bao bị missing value ở tháng thứ 1 (apru, call, data). Missing value xảy ra tương đồng theo các tháng ở 3 cột apru, data, call. Do đó, nếu thuê bao bị missing value ở apru_2 thì call_2 và data_2 cũng bị missing values và tương tự cho các tháng khác |

Ta có phương án  khác drop các thuê bao khi bị missing value. Ta có tiêu chí drop theo hàng:

- Nếu không đủ thông tin để mô tả hay phán đoán một mẫu (missing trên 80% số cột. Ví dụ: Người dùng này có 13/15 cột bị missing values thì bỏ đi người dùng này. Ta chấp nhận ngưỡng 80% vì trong dataset có những user chỉ mới bắt đầu sử dụng dịch vụ và được ghi nhận ở tháng thứ 5, tức khi đó 12/15 cột bị missing value)

Sau khi loại bỏ, ta còn 330857 mẫu thuê bao.

## Phần 3: Thay thế các missing value với giá trị Max của các cột ngay_goi_tong_dai_gan_nhat, ngay_dang_ky_goi_gan_nhat, ngay_nap_tien_gan_nhat

Max values thể hiện xu hướng ít gắn bó trong ngữ cảnh này, thời gian càng dài chứng tỏ thuê bao không có nhiều quan tâm đến gói/tổng đài/nạp tiền. Phương án này giúp cho ta khi visualize phân biệt được tương quan giữa các thuê bao bị khóa và mở khóa theo sự tương tác gần đây nhất với dịch vụ.

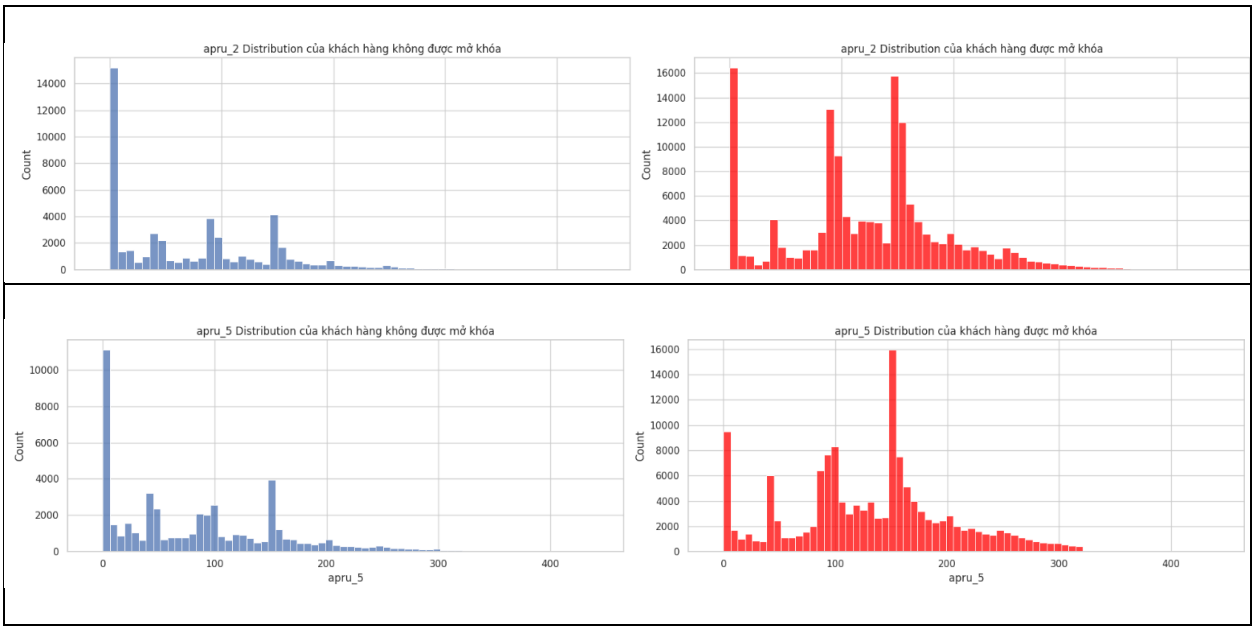
## Phần 4: Scale lại apru

Vì apru có đơn vị là đồng, khi visualize sẽ gặp phải ngoại lai nhiều làm cho biểu đồ bị thưa ra và nhỏ đi. Ta cần scale lại để việc visualize được tốt hơn

# Phần: Phân tích EDA.

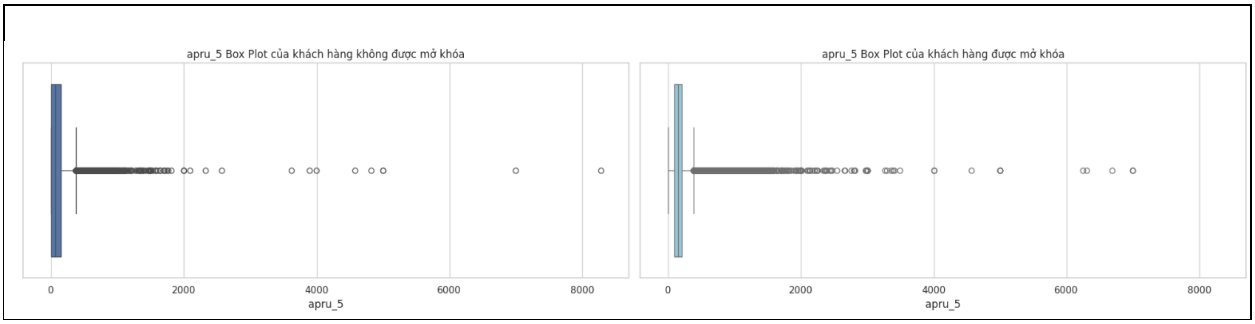
Trích chọn đặc trưng cơ bản của nhóm 2 nhóm khách hàng

## Phần 1: Doanh thu chủ yếu ở 2 nhóm

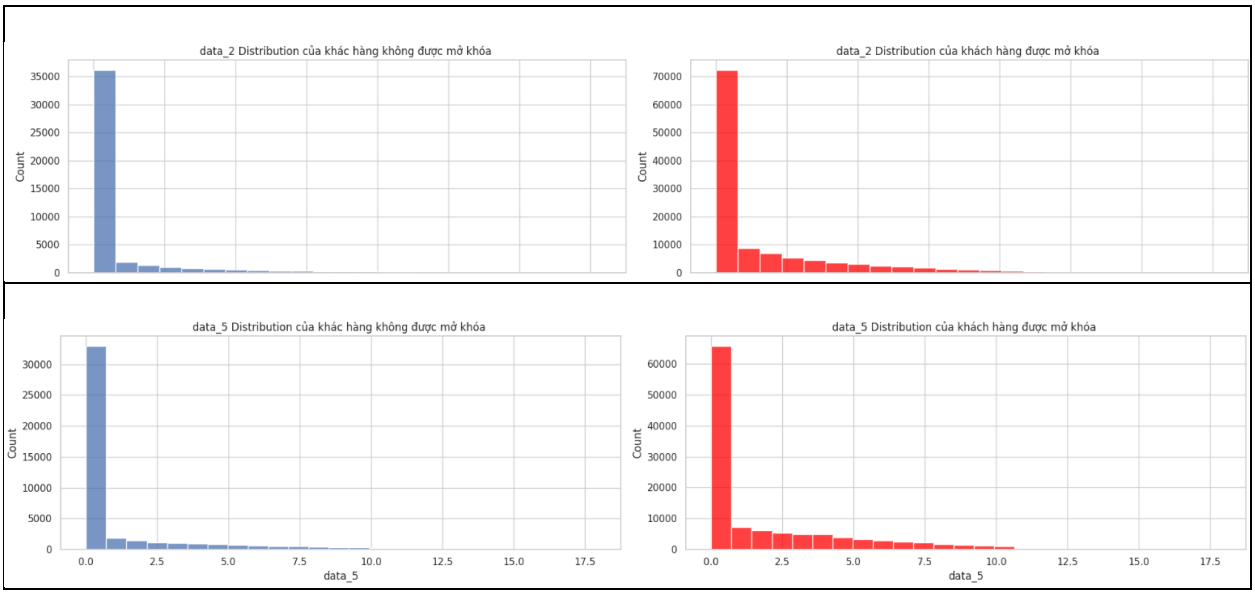



Ta thấy rằng doanh thu chủ yếu tập trung ở mức 50 nghìn, 100 nghìn, và nhiều nhất 150 nghìn. Trong đó, ở nhóm được mở khóa thì số lượng ở 3 mốc này tập trung nhiều hơn so với nhóm không được mở khóa.

Ngoài ra, ngoại lai xảy ra rất nhiều trong các tháng, tiêu biểu là tháng thứ 5 có khách hàng có tận dung thu là 8 triệu nhưng lại bị khóa thuê bao

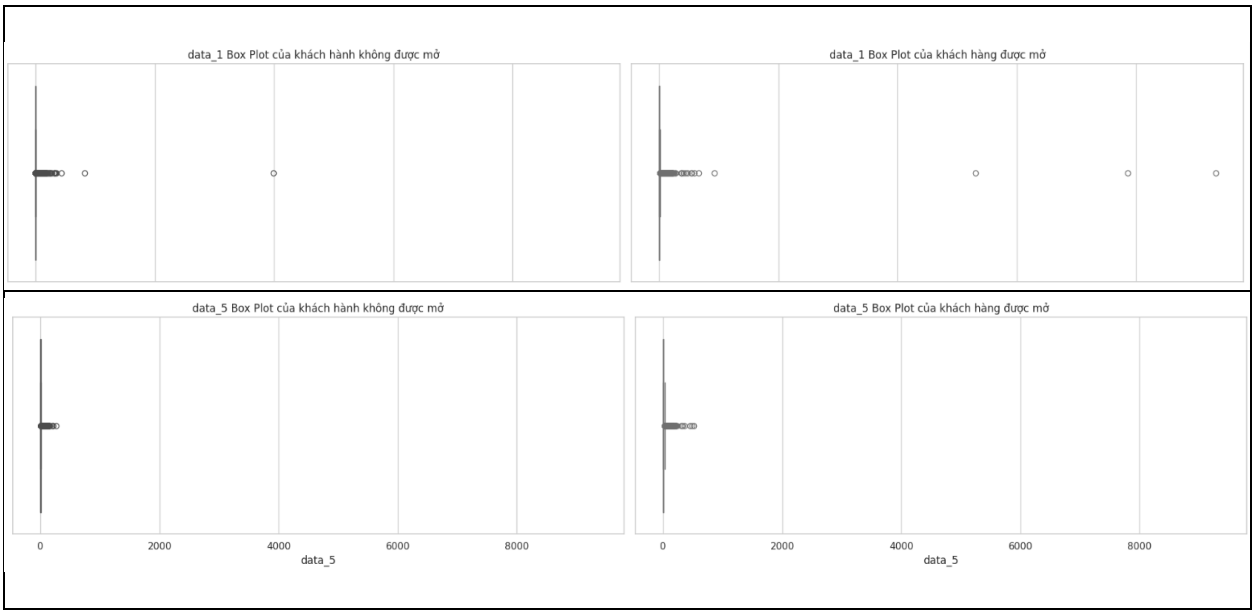


## Phần 2: Lượng data sử dụng của các thuê bao

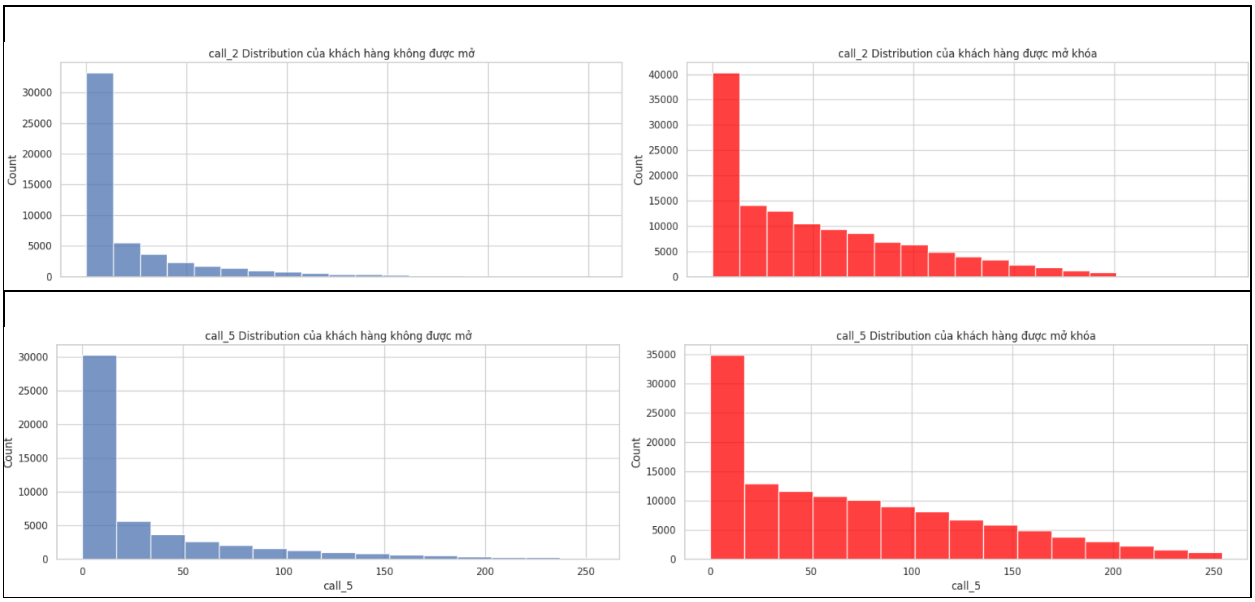



Phần lớn lượng dữ liệu sử dụng ít, nhóm được mở khóa thì lượng người sử dụng dữ liệu nhiều hơn ở các mốc.

Ngoài ra, ngoại lai cũng xuất hiện nhiều, cá biệt có thuê bao dù lợi nhuận trong 2 tháng tương đương nhau nhưng lượng data trong 2 tháng lại chênh lệch rất nhiều (8kGb), điều này có thể là do sai sót khi nhập liệu. Khi phán đoán cho mô hình cần xử lý thêm.



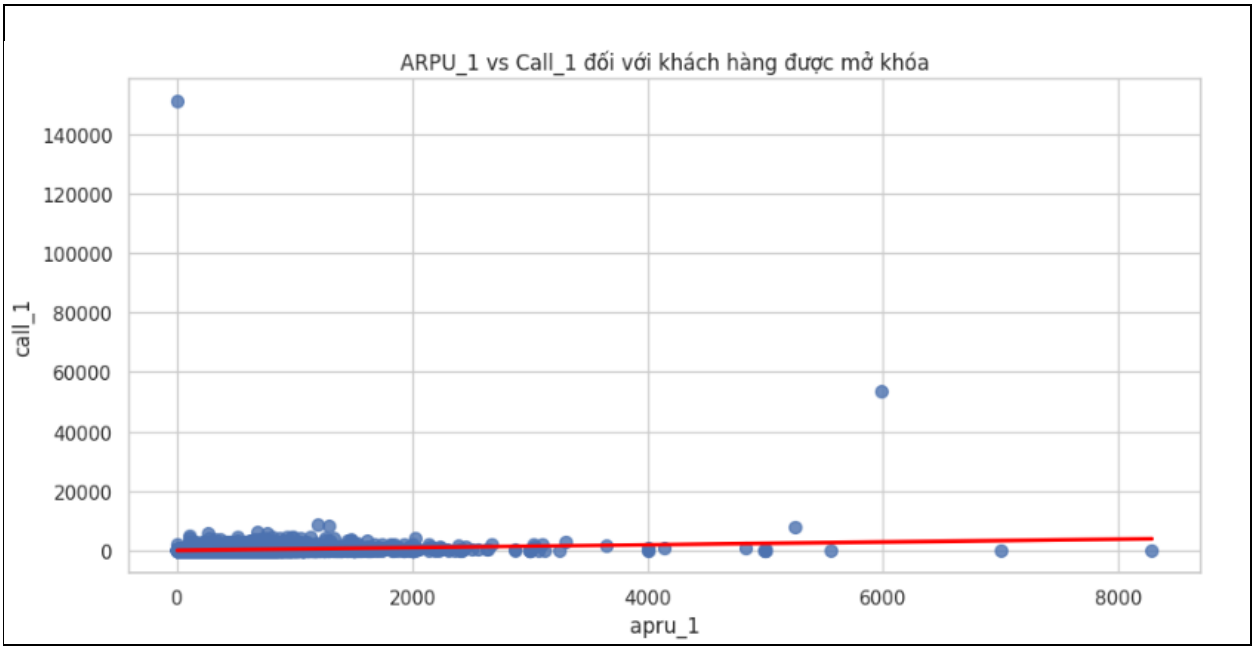
## Phần 3: Số lượng cuộc gọi

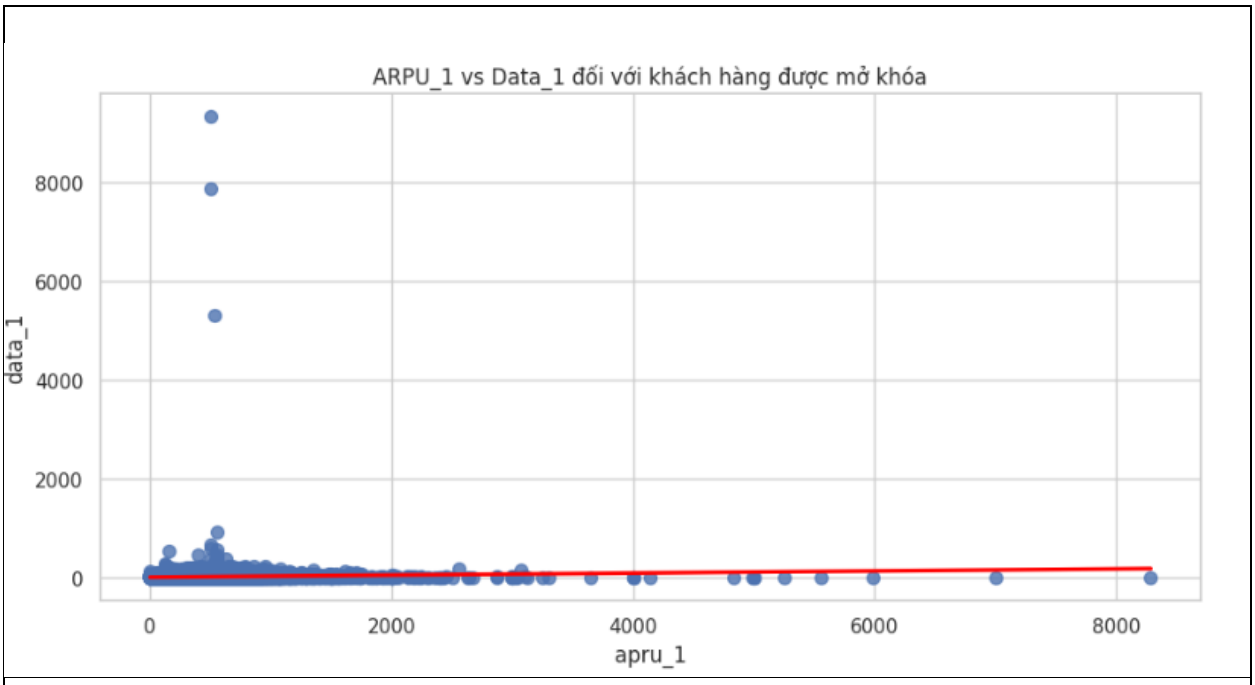


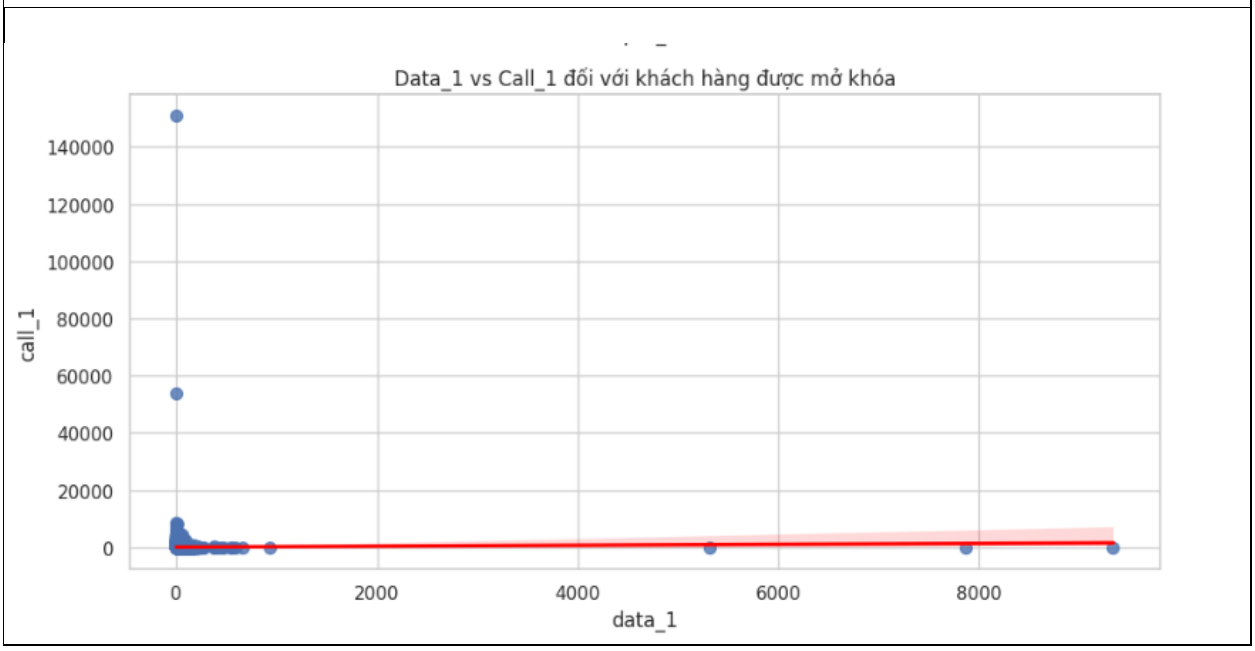

Tình trạng ngoại lai xảy ra tương tự như ở đặc trưng data.

## Phần 4: Xem xét tương quan giữa các cặp data, arpu và call

Các đặc trưng này không có sự tương quan mạnh với nhau ở  nhóm thuê bao được mở khóa (tương tự với nhóm không được mở). Dù cho lợi nhuận cao nhưng số lượng cuộc gọi lại không có nhiều ảnh hưởng, hay dữ liệu dử dụng. Ngoài ra, có một số trường hợp ngoại lai đặc biệt cần lưu ý như lượng data =0 nhưng số lượng cuộc gọi lại rất nhiều.

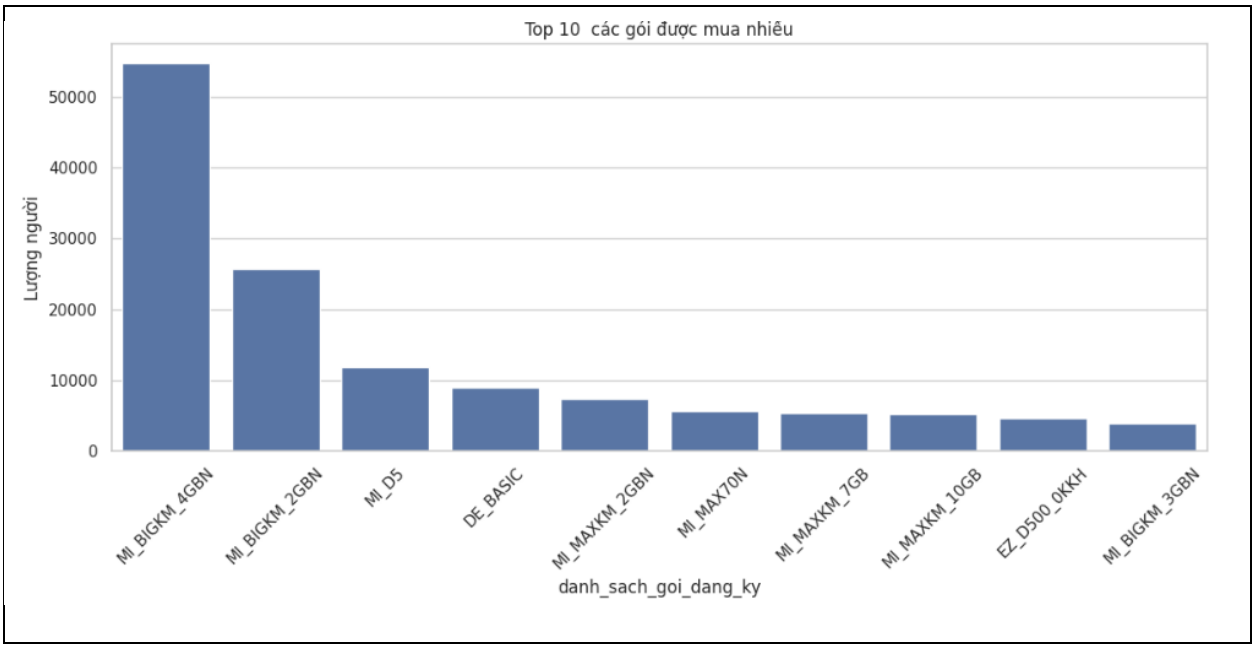




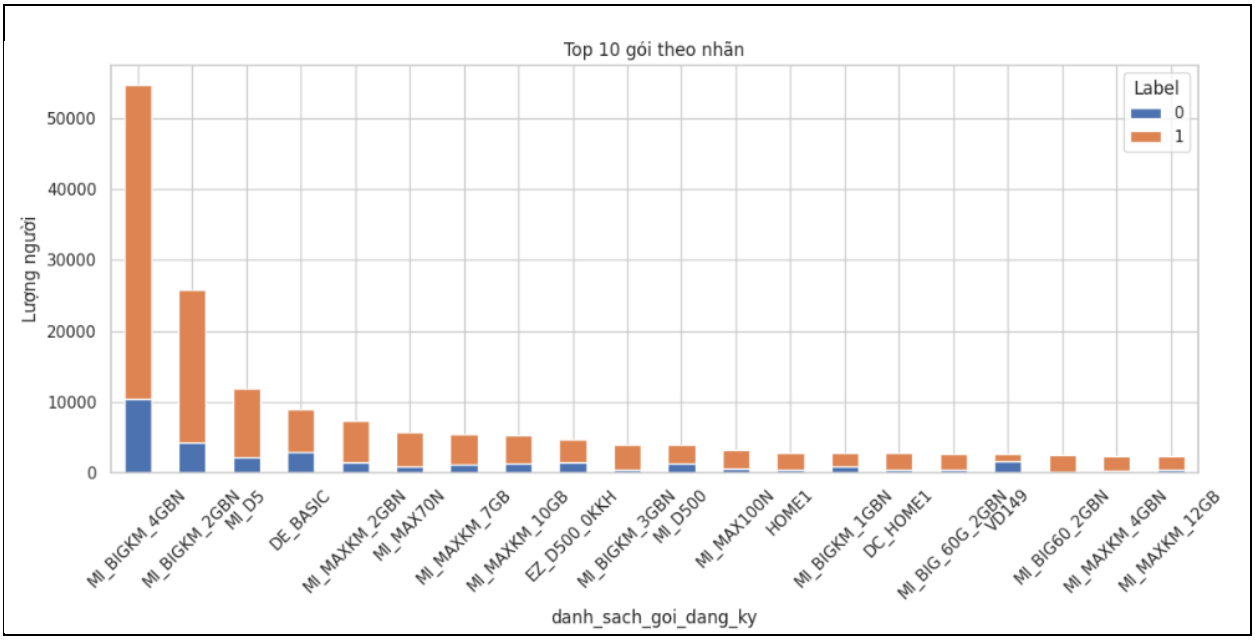




## Phần 5: Việc đăng ký gói của khách hàng

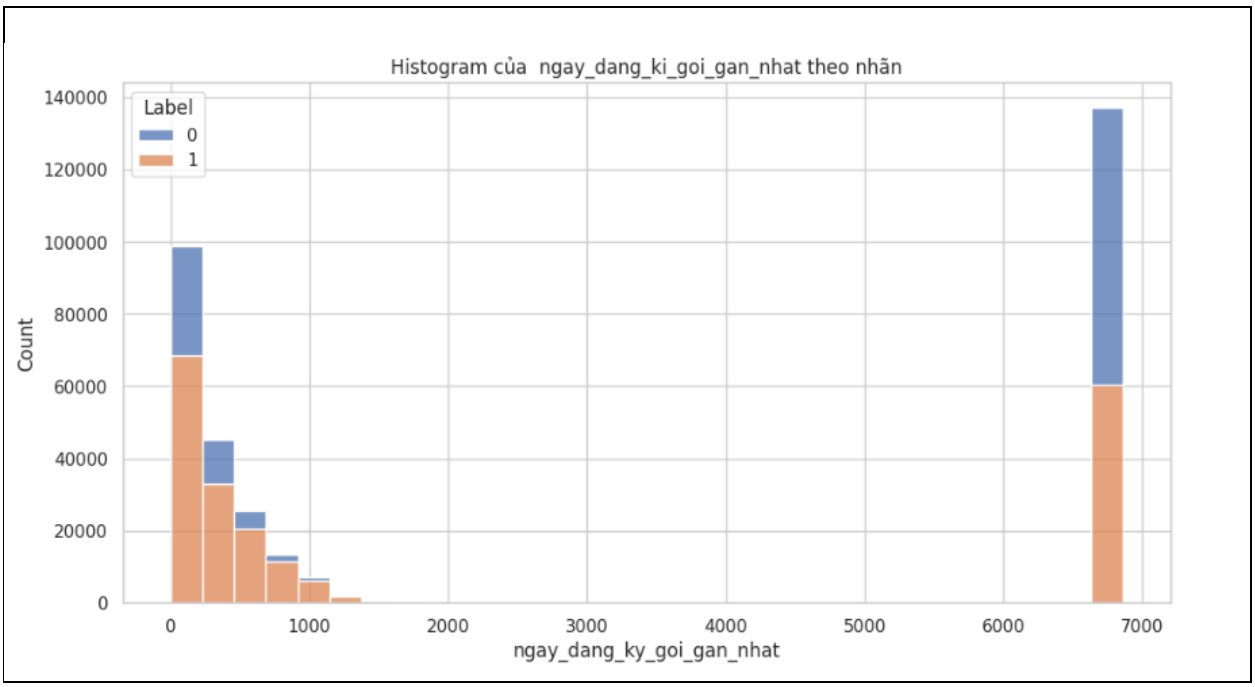
Có 704 gói được cung cấp. Trong đó MI_BIGHM_4GBN vàMI_BIGKM_2GBN được dùng nhiều nhất



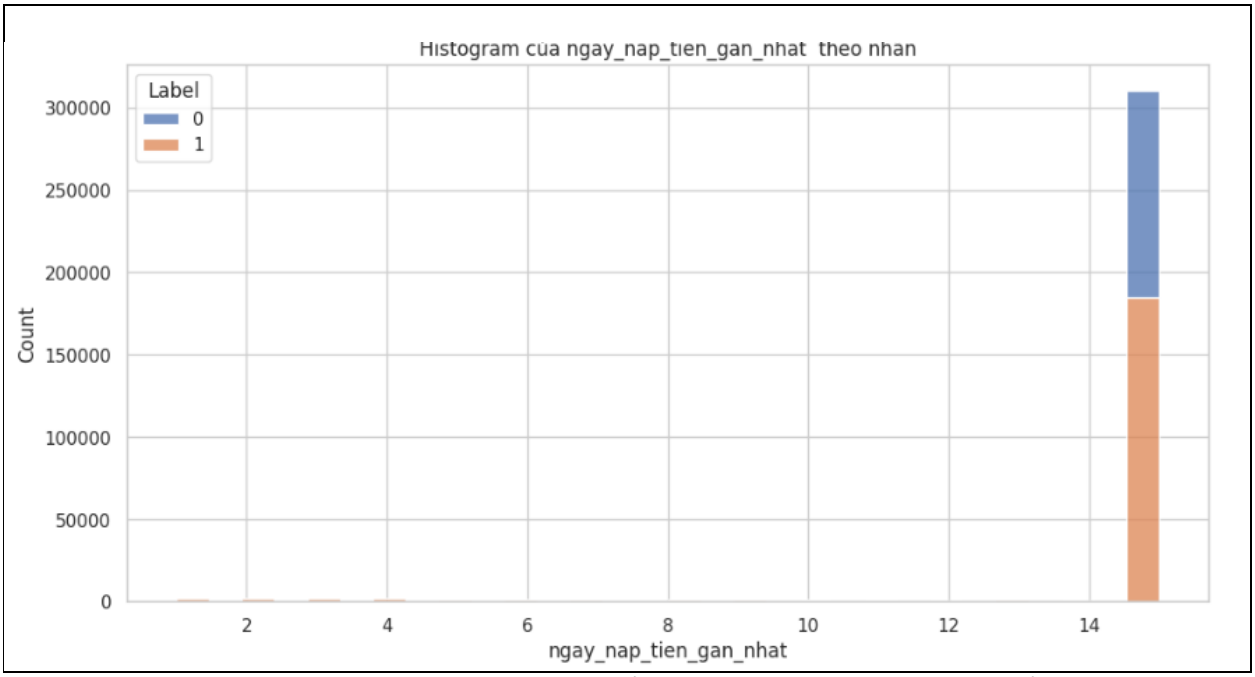

Đối với 2 nhóm thuê bao thì có phản ánh xu hướng ở trên các gói: đăng ký gói có tương quan đến việc mở khóa thuê bao. Tuy nhiên, vì phần lớn gói nào cũng có xu hướng như vậy nên ta không thể trích chọn hết các gói (hơn 700 gói) là những đặc trưng của mô hình.



## Phần 6: Tương quan giữa ngày đăng ký gói gần nhất, ngày nạp tiền gần nhất đến từng nhãn

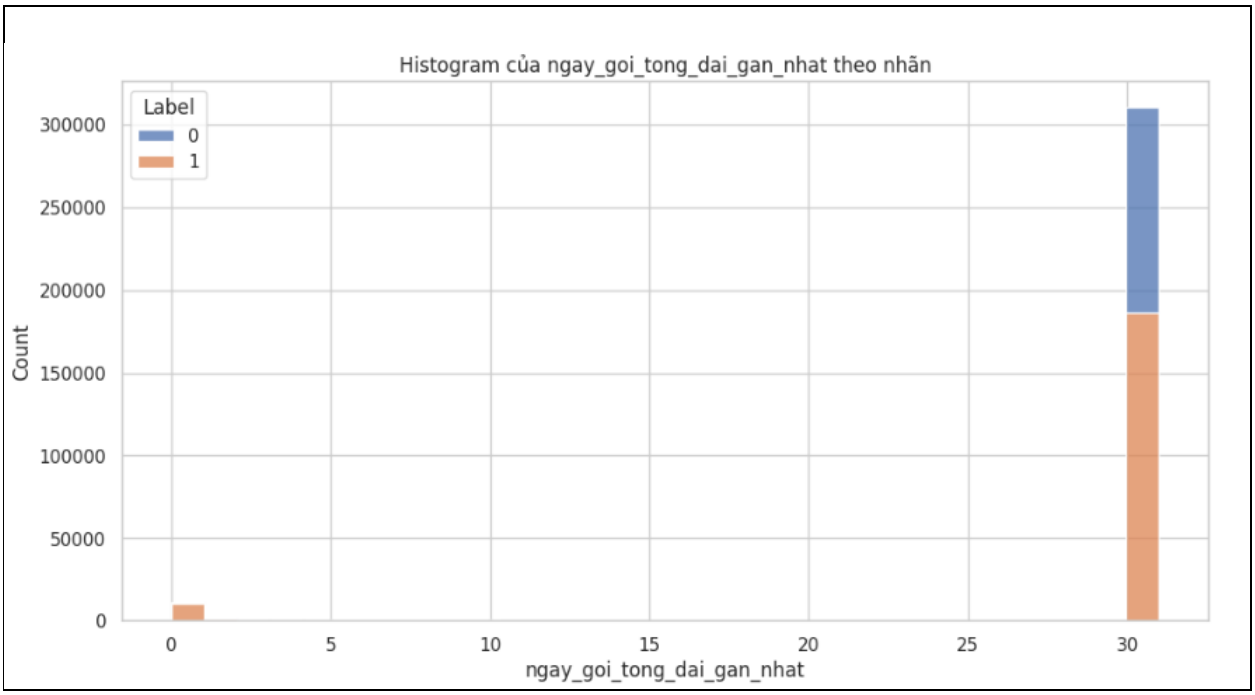



Do đã fill với max value, ta thấy rằng phần lớn khách hàng được mở khóa thuê bao thì có xu hướng đăng ký gói gần với hiện tại hơn, và ngược lại, phần lớn nhóm không mở khóa thì ít đăng ký gói (ở ngữ cảnh này là đã đăng ký gói từ rất lâu rồi).




Đối với ngày nạp tiền gần nhất, ở 2 nhóm không thể hiện sự khác biệt, nên ta có thể loại bỏ trong mô hình về sau (và thực tế là để tránh data leakage do ta phán đoán được mối liên hệ giữa ngày nạp tiền

gần nhất và nhãn của mô hình). Tương tự, ta không thấy sự khác biệt ở 2 nhóm thuê bao đối với đặc trưng ngày gọi tổng đài gần nhất.




Vậy các đặc trưng hầu hết sẽ được giữ lại trừ danh sách gói đăng ký và ngày nạp tiền gần nhất và ngày gọi tổng đài gần nhất.

# Phần: Huấn luyện mô hình và đánh giá

## Phần 1: Xử lý ngoại lai

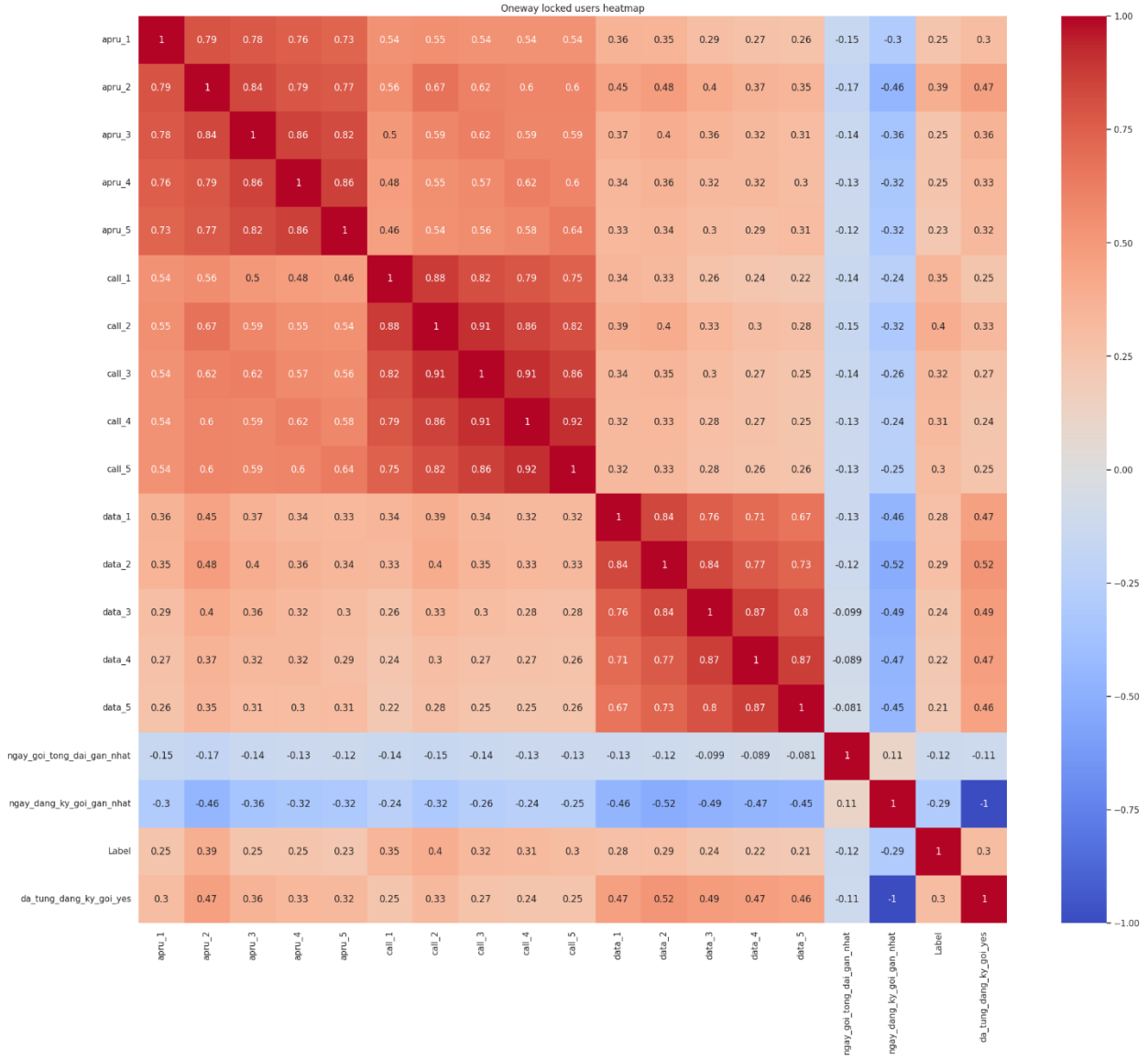
Ta đã nhận thấy những ngoại lai đáng ngại ở bước EDA trên. Ta sẽ sử dụng capping method để thay thế những giá trị ngoại lai bằng giá trị giới hạn (tứ phân vị thứ 3 Q3).

## Phần 2: Filling với không, và median ở các cột Arpu, call, data

Ta có các đặc trưng về ARPU, Data, Call của 5 tháng gần nhất tính tới ngày bị khóa. Do đó, những nan values xuất hiện trong này có thể là việc thuê bao bị khóa nên hệ thống không ghi nhận cụ thể. Do đó, khi huấn luyện, ta sẽ fill bằng 0. Ngoài ra, ta sẽ thử cách fill bằng median do đặc trưng phân phối của dữ liệu.

Với 2 phương pháp filling bằng không và bằng median thì kết quả mang lại ở phương pháp median tốt hơn trên các mô hình.

## Phần 3: Kiểm tra tương quan bằng heatmap




Để ý đã từng đăng ký gói và ngày đăng ký gói gần nhất có tương quan = -1, điều này thực tế không đúng với thực tế đã biết, ta sẽ cần thử nghiệm để bỏ đi cột nào.

## Phần 3: Huấn luyện với XGBOOST và RF

| | Accuracy | Recall | Precision | F1 | AUC-ROC |
|---|---|---|---|---|---|
| Random forest | 0.7984 | 0.8060 | 0.8843 | 0.8433 | 0.7731 |
| XGBOOST | 0.7955 | 0.8089 | 0.8729 | 0.8397 | 0.7727 |

Random forest cho kết quả tốt nhất trên các độ đo mong muốn như f1 và auc-roc.

## Phần 4: Tunning Random forest

Random Forest Parameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}

| | Accuracy | Recall | Precision | F1 | AUC-ROC |
|---|---|---|---|---|---|
| Random forest | 0.8013 | 0.8073 | 0.8882 | 0.8458 | 0.7757 |

## Phần: Phân tích lỗi.

Phần 1: Ta so sánh  độ quan trọng đặc trưng của mô hình đã huấn luyện với trung bình độ quan trọng đặc trưng trên các mẫu bị sai (false negative, tức thực tế cần mở khóa thuê bao nhưng mô hình dự đoán là vẫn khóa).

| | Feature | Importance | Mean Importance for Misclassified |
|---|---|---|---|
| 0 | call_2 | 0.104704 | 0.541171 |
| 1 | call_3 | 0.095092 | 1.389038 |
| 2 | apru_2 | 0.088763 | 3.156503 |
| 3 | apru_3 | 0.074741 | 5.037867 |
| 4 | call_4 | 0.070088 | 1.440281 |
| 5 | call_1 | 0.069376 | 0.586855 |
| 6 | call_5 | 0.068057 | 1.918862 |
| 7 | ngay_dang_ky_goi_gan_nhat | 0.058285 | 245.580018 |
| 8 | apru_4 | 0.055969 | 4.020106 |
| 9 | apru_5 | 0.052991 | 4.126886 |
| 10 | data_1 | 0.050766 | 0.022779 |
| 11 | data_3 | 0.049180 | 0.062198 |
| 12 | apru_1 | 0.045511 | 2.939910 |
| 13 | data_5 | 0.039145 | 0.081822 |
| 14 | data_4 | 0.036383 | 0.061316 |
| 15 | data_2 | 0.035102 | 0.021759 |
| 16 | da_tung_dang_ky_goi_yes | 0.005848 | 0.002345 |

Như vậy, trên các mẫu bị sai là do ngày đăng ký gói gần nhất dominate các đặc trưng còn lại. Kết hợp với thông tin đã biết từ heatmap, ta sẽ thử drop đi cột này và huấn luyện lại.

Ngoài ra các đặc trưng arpu cũng có sai lệch đáng kể. Ta sẽ xem xét phân phối của các mẫu bị sai (false negative) này với phân phối của 2 nhóm thuê bao.

Dự đoán sai (Label=1, Predict=0)

| | apru_1 | apru_2 | apru_3 | apru_4 | apru_5 |
|---|---|---|---|---|---|
| count | 7392.000000 | 7392.000000 | 7392.000000 | 7392.000000 | 7392.000000 |
| mean | 63.730623 | 35.024495 | 67.400450 | 72.196045 | 78.834030 |
| std | 75.405386 | 59.683011 | 67.559787 | 71.408508 | 72.865780 |
| min | -0.100000 | 0.000000 | -0.000400 | 0.000000 | 0.000000 |
| 25% | 0.000000 | 0.000000 | 1.098075 | 1.098125 | 4.980050 |
| 50% | 41.666667 | 0.596400 | 52.249050 | 59.333333 | 85.847550 |
| 75% | 120.000100 | 49.999325 | 110.691900 | 120.640300 | 124.166667 |
| max | 443.170887 | 384.426250 | 376.165625 | 386.806900 | 396.500500 |

Label=1

| | apru_1 | apru_2 | apru_3 | apru_4 | apru_5 |
|---|---|---|---|---|---|
| count | 203367.000000 | 203367.000000 | 203367.000000 | 203367.000000 | 203367.000000 |
| mean | 151.154810 | 133.201212 | 143.119691 | 150.020032 | 152.713205 |
| std | 111.596665 | 107.656620 | 103.285339 | 101.105755 | 104.210524 |
| min | -0.100000 | 0.000000 | -9.000000 | -0.000500 | -0.000750 |
| 25% | 90.349800 | 44.999999 | 83.999300 | 90.290200 | 90.399100 |
| 50% | 130.984700 | 126.346395 | 129.699900 | 130.742304 | 131.448900 |
| 75% | 188.999800 | 178.519313 | 188.999800 | 190.941850 | 195.352300 |
| max | 443.170887 | 384.426250 | 376.165625 | 386.806900 | 396.500500 |

Label=0

| | apru_1 | apru_2 | apru_3 | apru_4 | apru_5 |
|---|---|---|---|---|---|
| count | 127490.000000 | 127490.000000 | 127490.000000 | 127490.000000 | 127490.000000 |
| mean | 95.598085 | 47.768728 | 92.637602 | 100.101944 | 105.984022 |
| std | 87.576998 | 81.814539 | 75.746462 | 76.148484 | 76.171087 |
| min | -0.100000 | -0.000500 | -0.000500 | -0.000500 | -0.000550 |
| 25% | 0.000000 | 0.000000 | 24.500300 | 41.666667 | 48.999500 |
| 50% | 120.000100 | 0.000000 | 110.691900 | 120.640300 | 124.166667 |
| 75% | 120.000100 | 88.999900 | 110.691900 | 120.640300 | 124.166667 |
| max | 443.170887 | 384.426250 | 376.165625 | 386.806900 | 396.500500 |

Rõ ràng, phân phối của nhóm dự đoán sai (xét theo arpu) sát với nhóm thuê bao bị khóa hơn về trung bình và độ lệch chuẩn. Như vậy, trường hợp false postive phần lớn là thuê bao có sinh lời ít, ta có thể không cần quan tâm đến nhóm thuê bao bị phân loại sai này vì họ không phải khách hàng tiềm năng xét theo lợi nhuận.

Phần 2: Ta so sánh  độ quan trọng đặc trưng của mô hình đã huấn luyện với trung bình độ quan trọng đặc trưng trên các mẫu bị sai (false positive, tức thực tế không cần mở khóa thuê bao nhưng mô hình dự đoán là mở khóa).

Dự đoán sai (Label=0, Predict=1)

| | apru_1 | apru_2 | apru_3 | apru_4 | apru_5 |
|---|---|---|---|---|---|
| count | 13111.000000 | 13111.000000 | 13111.000000 | 13111.000000 | 13111.000000 |
| mean | 129.889933 | 107.457254 | 117.424454 | 127.030028 | 131.353989 |
| std | 110.063371 | 103.163660 | 100.704413 | 96.845431 | 97.705338 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | -0.000550 |
| 25% | 57.244550 | 0.292000 | 41.655950 | 54.168800 | 66.104250 |
| 50% | 120.000100 | 90.954000 | 101.999700 | 120.640300 | 124.166667 |
| 75% | 157.801050 | 152.089900 | 157.000800 | 159.001150 | 159.918000 |
| max | 443.170887 | 384.426250 | 376.165625 | 386.806900 | 396.500500 |

Label=1

| | apru_1 | apru_2 | apru_3 | apru_4 | apru_5 |
|---|---|---|---|---|---|
| count | 203367.000000 | 203367.000000 | 203367.000000 | 203367.000000 | 203367.000000 |
| mean | 151.154810 | 133.201212 | 143.119691 | 150.020032 | 152.713205 |
| std | 111.596665 | 107.656620 | 103.285339 | 101.105755 | 104.210524 |
| min | -0.100000 | 0.000000 | -9.000000 | -0.000500 | -0.000750 |
| 25% | 90.349800 | 44.999999 | 83.999300 | 90.290200 | 90.399100 |
| 50% | 130.984700 | 126.346395 | 129.699900 | 130.742304 | 131.448900 |
| 75% | 188.999800 | 178.519313 | 188.999800 | 190.941850 | 195.352300 |
| max | 443.170887 | 384.426250 | 376.165625 | 386.806900 | 396.500500 |

Label=0

| | apru_1 | apru_2 | apru_3 | apru_4 | apru_5 |
|---|---|---|---|---|---|
| count | 127490.000000 | 127490.000000 | 127490.000000 | 127490.000000 | 127490.000000 |
| mean | 95.598085 | 47.768728 | 92.637602 | 100.101944 | 105.984022 |
| std | 87.576998 | 81.814539 | 75.746462 | 76.148484 | 76.171087 |
| min | -0.100000 | -0.000500 | -0.000500 | -0.000500 | -0.000550 |
| 25% | 0.000000 | 0.000000 | 24.500300 | 41.666667 | 48.999500 |
| 50% | 120.000100 | 0.000000 | 110.691900 | 120.640300 | 124.166667 |
| 75% | 120.000100 | 88.999900 | 110.691900 | 120.640300 | 124.166667 |
| max | 443.170887 | 384.426250 | 376.165625 | 386.806900 | 396.500500 |

Trường hợp này thì ngược lại, khi so sánh với phân phối về arpu đối với 2 nhóm, mô hình cho rằng nhóm này sinh lời tốt nên dự đoán là mở khóa.

Kết luận: Mô hình dự đoán khá tốt với mục tiêu tìm kiếm những thuê bao có lợi nhuận tốt để mở khóa. Thực tế, việc gán nhãn còn đơn giản, chưa xét đến trường hợp người dùng không chủ động nhắn mở thuê bao,... Với những thuê bao sinh lời thấp, khách hàng có thể chủ động nhắn gửi để mở khóa hoặc dịch vụ dành cho họ mức ưu tiên thấp để mở khóa.