

**TẬP ĐOÀN VIỄN THÔNG QUÂN ĐỘI**

**TÀI LIỆU XÂY DỰNG MODEL**

**MỞ RỘNG ĐỘ PHỦ DỰ ĐOÁN TẬP 3K3D KHÔNG ĐẠT VERSION 1**

**Mã hiệu dự án:**

**Mã hiệu tài liệu:**

Hà Nội, 22/02/2021

**MỤC LỤC**

[**1**](#_gjdgxs) **Mục tiêu 3**

[**1.1**](#_30j0zll) **Đầu vào của model 3**

[**1.2**](#_1fob9te) **Đầu ra của model 4**

[**2**](#_3znysh7) **Phương pháp 5**

[**2.1**](#_2et92p0) **Định nghĩa class 5**

[**2.2**](#_tyjcwt) **Shortlist các feature quan trọng 5**

[**2.3**](#_3dy6vkm) **Tuning / improve mô hình 7**

[**3**](#_1t3h5sf) **Kết quả mô hình 7**

[**4**](#_17dp8vu) **Hạn chế và các phương án cải thiện 9**

[**4.1**](#_3rdcrjn) **Hạn chế 9**

[**4.2**](#_26in1rg) **Đề xuất cải thiện 9**

# Mục tiêu

Dựa vào thông tin và hành vi sử dụng dịch vụ viễn thông của các thuê bao của VTT để tạo ra model dự đoán thuê bao nào đang hoạt động và đạt 3k3d tại tháng N, sẽ có xu hướng không đạt 3k3d tại tháng n+1, đối tượng được dự báo thuộc phân khúc khách hàng mass

Kết quả dự đoán trên sẽ được sử dụng để triển khai các chương trình khuyến mại, để rồi offer cho các khách hàng này các gói cước hay các chương trình khuyến mại phù hợp, các chương trình tác động để thực hiện mục tiêu kép giúp hồi phục tiêu dùng của khách hàng và góp phần giữ chân khách hàng, ngăn chặn hành vi suy giảm nhu cầu sử dụng thiết bị và giảm tiêu dùng xuống dưới 3k dẫn tới trạng thái không đạt thực 3k3d trước khi thuê bao rơi vào trạng thái cắt hủy rời mạng, làm giảm tỷ lệ rời mạng.

Tập thuê bao cần dự đoán (customer base) phải thỏa mãn các điều kiện sau:

* Vẫn đang còn hoạt động tại tháng n (status = '2'),
* Is\_3k3d = 1 ( là thuê bao đạt 3k3d tại tháng n )
* Tuổi thuê bao >= 2
* Là thuê bao mobile
* Thuộc segment Mass

SLTB của customer base trong các tháng như sau:

| **month** | **count(1)** |
| --- | --- |
| 20201101 | 23016584 |
| 20201201 | 24032581 |
| 20210101 | 21383313 |

Ghi chú:

* month là tháng N ( tháng xét để lấy dữ liệu thuê bao )
  1. **Đầu vào của model**

Bảng dữ liệu (dataset) chứa các đặc trưng về hành vi sử dụng dịch vụ viễn thông và thông tin thuê bao được tổng hợp trong 6 tháng gần nhất (do tập customer base tenure >= 2 tháng tuổi, nên các feature chỉ lấy trong tháng N, N-1, N-2, N-3, N-4, N-5 để đảm bảo dữ liệu đầy đủ trên tất cả các thuê bao)

Các dữ liệu có trong dataset:

* Dữ liệu về các thuê bao: tariff plan của thuê bao.
* Dữ liệu hành vi sử dụng: lưu lượng thoại, tin nhắn, data, vas trong quá khứ.
* Dữ liệu tiêu dùng: doanh thu thoại, tin nhắn, data, vas trong quá khứ.
* Dữ liệu nạp thẻ: số lần nạp thẻ, mệnh giá, tổng doanh thu nạp thẻ trong quá khứ
* Dữ liệu tần suất sử dụng dịch vụ ( gói cước )

Cấu trúc của dataset:

* Bảng input của model sẽ có định dạng như sau:

| **year** | **month** | **sub\_id** | **other\_feature\_1** | **...** | **other\_feature\_X** |
| --- | --- | --- | --- | --- | --- |
| 2020 | 06 | A |  |  |  |
| 2020 | 07 | A |  |  |  |
| 2020 | 08 | A |  |  |  |
| 2020 | 09 | B |  |  |  |
| 2020 | 10 | B |  |  |  |
| 2020 | 11 | B |  |  |  |

* Khi thực hiện train và predict, bộ các trường định danh [year, month, sub\_id, isdn, partition] sẽ không được sử dụng. Mỗi 1 row sẽ tương ứng với 1 data point (không có mối liên hệ giữa các row, kể cả các row của cùng 1 thuê bao tại các tháng khác nhau).
  1. **Đầu ra của model**

Output của model là với mỗi một {thuê bao}, dự đoán propensity score của thuê bao sẽ không phát sinh lưu lượng trong cả 2 tháng tiếp theo (N+1 và N+2).

Bảng output của model sẽ có định dạng như sau

| **sub\_id** | **other\_feature\_1** | **...** | **other\_feature\_X** | **score** |
| --- | --- | --- | --- | --- |
| A |  |  |  | 0.3 |
| A |  |  |  | 0.2 |
| A |  |  |  | 0.1 |
| B |  |  |  | 0.45 |
| B |  |  |  | 0.40 |
| B |  |  |  | 0.15 |

Tuy nhiên, để cho ngắn gọn, khi lưu kết quả predict xuống HDFS (trong bước infer), chỉ lưu các trường tối thiểu

| **sub\_id** | **Pred** | **sub\_id** | **score** |
| --- | --- | --- | --- |
| 1 | 0 | A | 0.3 |
| 3 | 1 | A | 0.2 |
| 4 | 0 | A | 0.1 |
| 5 | 1 | B | 0.45 |
| 6 | 0 | B | 0.40 |
| 7 | 1 | B | 0.15 |

# Phương pháp

* 1. **Định nghĩa class**

• Tập base:

→ Segment : mass

→là thuê bao mobile

→đang hoạt động tại tháng n

→ đạt 3k3d ( is\_3k3d = 1)

→ tuổi thuê bao >= 2

• Nhãn : không đạt 3k3d tại tháng n+1

(is\_3k3d = 0)

Và còn hoạt động

Trong đó

t\_org\_charge\_n: là tiêu dùng tháng n

t\_org\_charge\_n1: là tiêu dùng gốc tháng n + 1

t\_org\_charge\_n2: là tiêu dùng gốc tháng n + 2

* Trong dataset có 1 trường class:
  + **class**:
    - **== 1**: thuê bao nằm trong customer base tháng N và không đạt 3k3d trong tháng N+1
    - **== 0**: các thuê bao còn lại trong customer base.
* Số lượng thuê bao class 1 trong customer base hàng tháng (event rate):

| **Tháng n** | **SLTB class 0** | **Tỉ lệ class 0** | **SLTB class 1** | **Tỉ lệ class 1** | **Tổng SLTB** |
| --- | --- | --- | --- | --- | --- |
| 20201101 | 20520314 | 89.15% | 2496270 | 10.85% | 23016584 |
| 20201201 | 21119947 | 87.88% | 2912634 | 12.11% | 24032581 |
| 20211201 |  |  |  |  | 21383313 |

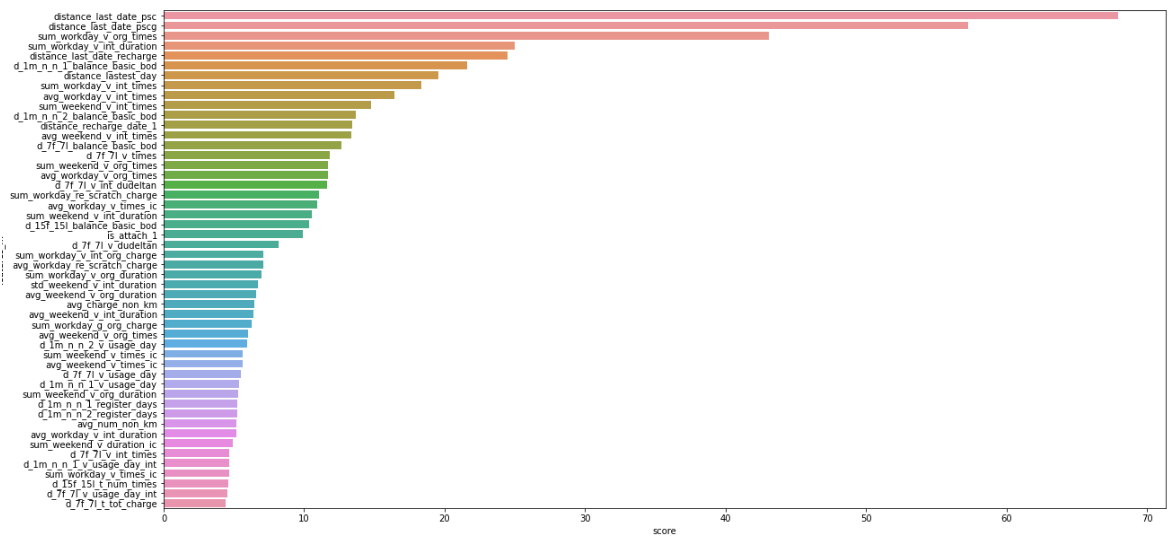
* Tỉ lệ class 1 tương đương khoảng 11=> 12% tùy từng tháng, tuy tỉ lệ này thuộc dạng imbalanced, nhưng SLTB class 1 tương đối nhiều (2.5-3 triệu thuê bao / tháng) nên có thể không cần áp dụng các hình thức up/down-sampling.
  1. **Shortlist các feature quan trọng**
* Do số lượng feature ban đầu rất lớn, 1337 feature, nên cần thực hiện shortlist những feature quan trọng nhất để đưa vào mô hình training. Việc shortlist feature có 2 lợi ích:
  + Giảm thời gian chạy của model, đặc biệt trong quá trình tiền xử lý dữ liệu và training
  + Giảm khả năng overfit của model
* Phương pháp shortlist thực hiện bằng cách build 1 model đơn giản và có khả năng đánh trọng số cho độ quan trọng (importances) của từng feature. Thông thường sử dụng các model tree-based. Trong bài toán này ta sử dụng 2 model đơn giản:
  + RandomForestClassifier với numTrees = 32
  + GBTClassifier với maxDepth = 6
* Trước khi đưa dữ liệu vào model trên, thực hiện các bước tiền xử lý dữ liệu như sau:
  + Bổ sung thêm các feature chưa có trong bảng c360 của Mytel, từ các bảng features bổ sung:
    - fe\_viettel\_synthesis\_m\_vs1\_${F\_DATE\_OF\_MONTH\_N}
    - f\_vtt\_pre\_features\_list
    - fe\_viettel\_categorical\_n\_${F\_DATE\_OF\_MONTH\_N}
    - testdb.c360\_v1\_9

Ghi chú : ${F\_DATE\_OF\_MONTH\_N} là tháng xét tại thời điểm lấy dữ liệu

* + Split data set thành 3 tập:
    - Train và valid: in-sample, chia theo tỉ lệ 80:20 randomly (tháng 11/2020)
    - Test: out-of-sample (tháng 12/2020)
  + Loại bỏ các feature có số trường null nhiều và các feature bị lỗi dữ liệu do bảng nguồn không cập nhật đầy đủ dữ liệu ở trạng thái thuê bao hoạt động
  + Loại bỏ các trường binary, do max\_depth của các tree model của Spark rất nông, chỉ có tối đa 30 depth, nên các trường binary sẽ "chiếm dụng" khoảng max depth này, trong khi thông tin của các trường binary lại bị hạn chế bởi chỉ có 2 giá trị 0/1 => model sẽ capture được lượng thông tin ít hơn so với các trường continuous.
  + Loại bỏ các trường không tin cậy, do người nhập dữ liệu: "customer\_age"
  + Loại bỏ các trường ít thông tin: "act\_status", customer\_age do tính chính xác dữ liệu không đảm bảo.
  + Tạm loại bỏ các trường có quá nhiều giá trị, khó sử dụng, chưa rõ cách tổng hợp (sale\_channel\_type, active\_channel\_type).
* Sau khi chạy thử với model đơn giản, kết quả như sau:
  + Performance (<http://10.60.129.158:9999/#/notebook/2FYEN7799>):

| **Tập thuê bao** | **AUROC** |
| --- | --- |
| train | 0.868 |
| valid | 0.861 |
| test | 0.855 |

* + Feature importances:



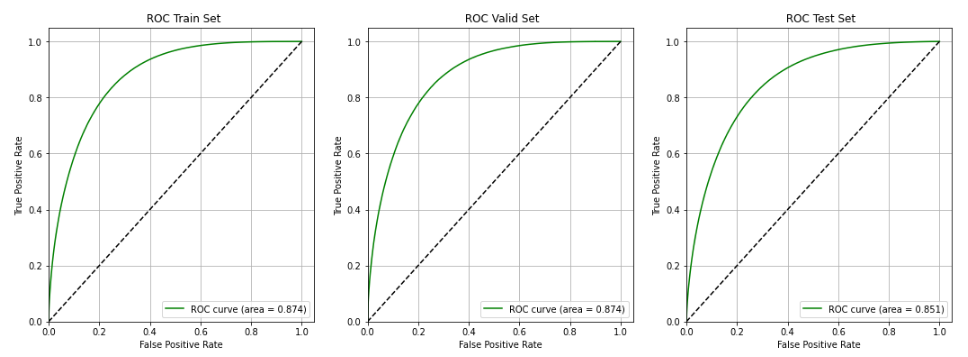
* Lựa chọn top 150 feature có độ quan trọng cao nhất, cho vào model training lại, được performance gần tương đương (<http://10.60.129.158:9999/#/notebook/2FYEN7799>):
* • select sub\_id from tmp\_model\_3k3d\_test\_result\_decile where decile >=1 and decile < 3;
* tương ứng tháng 11 dự báo tháng 12 phủ được 85~90% nhãn 1
* trong đó chia top decile thành 20 nhóm với prob từ cao xuống thấp tương ứng là nhóm decile 1 là nhóm có prob cao nhất và decile 20 là nhóm có prob thấp nhất
* Để lấy top ~ 30% theo lift gain thì chỉ cần lấy decile từ 1 đến < 3
* Tổng hợp số lượng thuê bao theo các nhóm decile:
* Feature

| **Tập thuê bao** | **AUROC** |
| --- | --- |
| train | 0.88786 |
| valid | 0.8734 |
| test | 0.8906 |

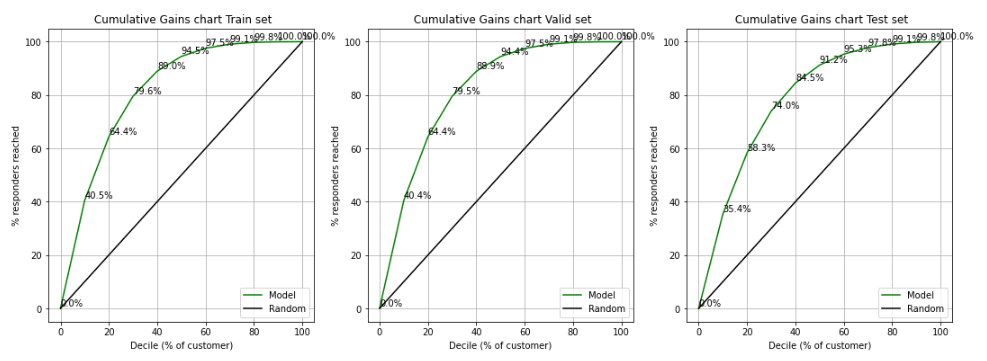
* Quyết định lựa chọn top 150 feature này vào danh sách short list: trong file "PL1 - churn feature shortlist.xlsx"
* Bảng feature important : model\_3k3d\_featureImportances
  1. **Tuning / improve mô hình**
* Do thời gian training với tập 150 feature nhanh hơn 1337 feature (30-45 phút / model), hơn nữa performance của model base đã đối tốt so với model base, nên hiện tại không subsample thành tập nhỏ hơn để tuning và improve trên scikit-learn, mà thực hiện tuning và improve luôn trên code Spark, ngoài ra có thể sử dụng thêm một số hàm thống kê và kỹ thuật phân lớp để đánh giá chất lượng feature model. Tuy nhiên chất lượng dữ liệu mytel không được đảm bảo do còn nhiều lỗi dữ liệu chưa đc clean sạch từ phía bảng nguồn. Nằm ở các trường quan trọng như num\_register\_day, balance\_basic ...
* Lựa chọn thuật toán RandomForest trong thư viện của Spark với bộ tham số sau là tối ưu:
  + numTrees= 60
  + minInstancesPerNode= 200
  + subsamplingRate= 1
  + maxDepth=30
  + maxBins=32

# Kết quả mô hình

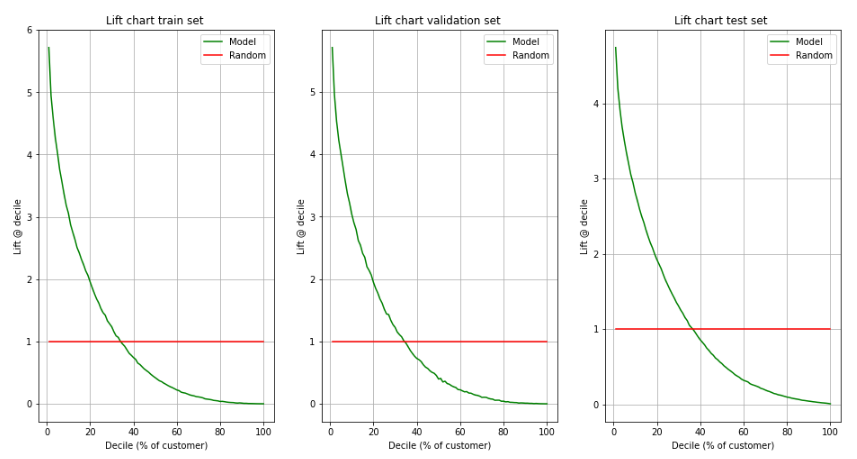
* ROC



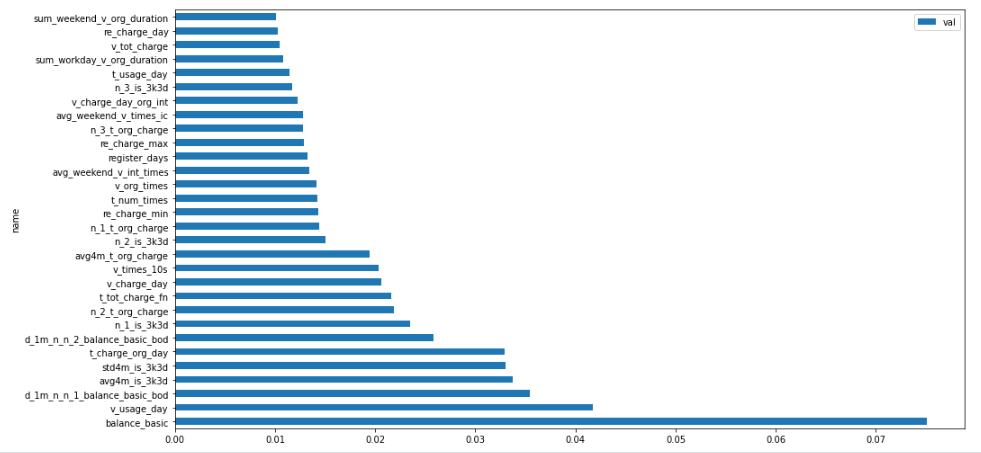
* Accumulative graphs (train valid test)



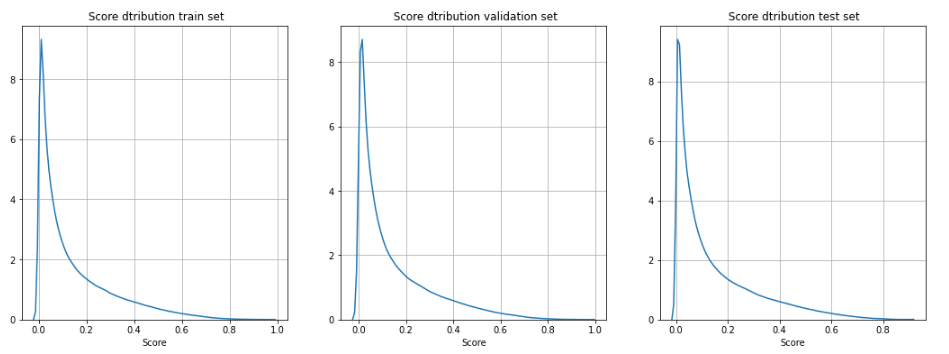
* Lift chart



* Feature importances:



* Score distribution:



# Hạn chế và các phương án cải thiện

* 1. **Hạn chế**
* Performance của model ở mức chấp nhận được, mặc dù chưa được cao vượt trội, do dữ liệu từ bảng c360 vẫn còn lỗi các trường nạp thẻ và bị duplicate bản ghi, chưa được fix lỗi đầy đủ. Nhìn vào top feature importances thì có thể thấy 2 loại feature có tính phân loại tốt nhất là t\_org\_charge, num\_register\_day và num\_usage\_day và balance\_basic, ngoài ra các điểm score feature khá sát nhau nên cho thấy hành vi suy giảm trong tập khá đa dạng và đều có ảnh hưởng đến kết quả model một cách sâu sắc.
* Điều này cho thấy target của model là phù hợp với logic phân loại thông thường bằng trực giác
  1. **Đề xuất cải thiện**
* Performance model đã cải thiện để nâng ROC trên tập test lên 0.9 và nhận định target khá dễ dự đoán nhưng ở nhóm score điểm thấp đặc biệt ở phân khúc khác hàng mass đã khá nhiều nhiễu do có nhiều thuê bao thuê bao kit tồn kênh, multisim, tiêu dùng ở mức thấp nên mức giảm nhỏ càng khó dự báo đc chính xác nên có thể để nâng được hiệu suất model ta nên cải thiện thêm các feature bổ sung và break nhỏ các nhóm model để dự đoán hành vi được chi tiết hơn.

Ví dụ, ta đứng ở tháng N+1, lấy dữ liệu tháng N để dự đoán các thuê bao sẽ suy giảm tiêu dùng tại tháng N+1 và N+2 so với tháng N. Tuy nhiên để model phân loại tốt ta cần dùng một số kỹ thuật thống kê phân lớp để tìm ra các nhóm có mức giảm tiêu dùng với hành vi cụ thể ( như giảm số ngày bật máy, giảm tiêu dùng thoại.....) hơn thì khi đó tổng hợp các nhóm model lại ta có được kết quả cuối cùng của nhóm suy giảm có hiệu suất phân loại chính xác hơn .

* Các phương án để cải thiện vấn đề mở rộng độ phủ:

• Sử dụng các thuật toán phân lớp K-mean, Kiểm định phân phối, Ước lượng phương sai....và các insight hành vi từ bên kinh doanh để đưa ra đc các nhóm cụ thể

* • Phương án 1 : chạy model nhóm điểm thấp  
  sử dụng phân cụm tập thuê bao và chạy theo nhóm nhỏ theo các hướng giảm tiêu dùng và số ngày bật máy, và nhãn các hành vi dự kiến là tương quan cao với khả năng không đạt 3k3d , kết hợp tăng số lượng feature  
  • Phương án 2: chạy model nhóm có thuê bao  
  → Đk : đạt 3k3d ( hoặc đạt điều kiện tháng n có tiêu dùng >3k và số ngày sử dụng >=3),tuổi thuê bao >=3 tại tháng n  
  nhãn giảm tiêu dùng xuống < 3k tháng n+1  
  • Phương án 3: chạy model nhóm giảm thời gian sử dụng  
  → Đk : đạt 3k3d ( hoặc đạt điều kiện tháng n có tiêu dùng >3k và số ngày sử dụng >=3), tuổi thuê bao >=3 tại tháng n  
  nhãn tiêu dùng xuống có t\_usage\_day < 3 ngày  
  • Phương án 4: Gaussian Mixture Model,GBT Classification Model, Naive Bayes Model, Multilayer Perceptron Classification  
  sử dụng phương pháp ensemble Learning phối hợp các thuật toán random forest và các thuật toán trên để lấy ra nhóm thuê bao xác suát cao nhất với nhóm model có điểm thấp đẩy ra .
* Các thuật toán sử dụng : random forest, Gaussian Mixture Model,GBT Classification Model, NaiveBayesModel, Multilayer Perceptron Classification  
  Hiện tại tháng này sẽ chạy thử nghiệm các phương án 1 2 3, sau khi chạy model con đã được phân tập sẽ lọc tập thuê bao thỏa mãn cả hai đk không đạt đk 3k3d tháng n+1 để dự báo  
  để xem số lượng thuê bao dự kiến được dự báo thêm, để đánh giá kết quả.  
  Phương án 4 đang khảo sát và thử nghiệm khi phương án 1 2 3 chưa đạt hiệu quả cao hoặc muốn phủ nhiều hơn
* đánh giá phần nhóm thuê bao kit tồn kênh,  để DS đánh giá tập này với nhóm model, vì khả năng nhóm này ko đạt 3k3d lớn.
  + Phân tích dự đoán của model trên 4 trường feature:
    - Thời điểm psc
    - Is\_3k3d
    - Balance\_basic
    - Num\_register\_day

Cùng các derived feature từ 4 feature này để có thể nhận biết sớm thuê bao có dấu hiệu không đạt thực có mức giảm tiêu dùng xuống dưới 3k và giảm nhu cầu tương tác sử dụng thiết bị xuống dưới 3 ngày/ tháng mạnh và tác động nhóm có tỷ lệ ko đạt thực nhiều nhất.