## Phần: Xây dựng bài toán.

### Phần 1: Bài toán

* Dự đoán giá trị doanh thu arpu của khách hàng chi tiêu ở mức cao.

Mục tiêu là sử dụng kết quả dự đoán của mô hình để phân nhóm những khách hàng doanh thu cao nhưng có xu hướng giảm tiêu dùng. Từ đó có đề xuất:

* Đối với Khách hàng có ARPU ổn định hoặc tăng, mô hình có thể đưa ra những gợi ý sau:

Ưu đãi gói cao cấp: Khách hàng có ARPU cao có thể quan tâm đến các gói cao cấp (ví dụ: giới hạn dữ liệu cao hơn, cuộc gọi quốc tế hoặc các dịch vụ đi kèm như phát trực tuyến).

Khuyến mãi bổ sung: Công ty có thể cung cấp các dịch vụ bổ sung (như gói chuyển vùng quốc tế hoặc giải trí) dựa trên thói quen chi tiêu của những khách hàng này và doanh thu dự đoán trong tương lai.

* Đối với Khách hàng có ARPU giảm đi:

Khuyến khích sự tương tác: Những khách hàng được dự đoán có ARPU giảm có thể nhận được các ưu đãi như dữ liệu thưởng hoặc số phút gọi miễn phí để tăng mức độ tương tác và mức sử dụng, từ đó tăng ARPU của họ.

### 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: Sử dụng dữ liệu của tháng 7,6,5,4
* Khách hàng doanh thu ở mức cao ở tháng thứ 7, trên 125 nghìn đồng.

### Phần 4: Chọn biến mục tiêu

* Doanh thu của tháng thứ 5 của từng khách hàng.

### Phần 5: Phương án đánh giá

<https://medium.com/@venugopal.adep/evaluation-metrics-in-regression-b6f3949ea448>

<https://medium.com/@tavishi.1402/part-1-performance-metrics-in-machine-learning-27cd9264e398>

<https://www.datasciencehorizon.com/data-science-with-python-performance-metrics-for-regression/>

<https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>

Mô hình dự đoán:

* Khả năng nhận biết được vùng giá trị doanh thu
* Doanh thu dự đoán sai lệch trung bình nằm trong dưới 50 nghìn đồng.

Bài toán áp dụng:

* Đề xuất 1: Cực đại hóa phát hiện số lượng thuê bao giảm doanh thu.
* Đề xuất 2: Số lượng thuê bao được phân nhóm gần đúng với phân nhóm thực

## Phần: Tiền xử lý

### Phần 1: Lọc các thuê bao nằm ở mức cao trở lên và đảm bảo điều kiện ít nhất 1 tháng hoạt động

Vì cần quan tâm đến dự báo doanh thu khách hàng trong tháng 8, ta phải biết thông tin hoạt động của người dúng trước đó trong vòng 4 tháng trở lại và có ít nhất 1 tháng hoạt động.

Ngoài ra, vì mục tiêu bài toán quan tâm đến người dùng có doanh thu cao, tức là trên 129 nghìn đồng.

Sau kkhi lọc, ta thu được 123741 thuê bao.

### Phần 2: Lọc bỏ các cột không cần thiết

Vì bối cảnh dữ liệu chỉ quan tâm thông tin sử dụng của 5 tháng trước đó, những thông tin sau đó có liên quan đến tháng 9 trở đi đều không phù hợp, nên ta sẽ lấy các cột arpu, data, call

### Phần 3: Lọc bỏ duplicate để tránh nhiễu

Có 3273 thuê bao trùng nhau về giá trị trong tất cả các cột đặc trưng. Sau khi lọc, ta còn 122042 thuê bao.

## Phần: EDA

### Phần 1: Phân phối của các đặc trưng, biến mục tiêu

|  |
| --- |
|  |
|  |

Doanh thu người dùng tập trung nhiều ở mức 200-250 nghìn đồng. Phân phối right-skewed. Doanh thu có xu hướng tăng và ổn định dần về phân phối.

|  |
| --- |
|  |
|  |

Phân phối lượng dữ liệu sử dụng thường tập trung nhiều ở 0-10GB một tháng. Lưu ý là tháng 7 lại ít ghi nhận thông tin người dùng nhất, và có ngoại lai lớn >5000GB

|  |
| --- |
|  |
|  |

Phân phối về số lượng cuộc gọi qua các tháng tương đồng nhau, và ở bin thứ 2 có xu hướng tăng dần (100-200 cuộc). Phân phối right-skewed. Giá trị ngoại lai không biến thiên nhiều như data

Ngoại lai xuất hiện rất nhiều ở các đặc trưng này, điều này yêu cầu ta phải xử lý chúng cho mô hình về sau:

|  |
| --- |
|  |
|  |
|  |

Đối với biến mục tiêu:

|  |
| --- |
|  |
|  |

Phân phối doanh thu tháng 8 lại có tương đồng với phân phối của tháng 4, như vậy có thể có những sự kiện theo Quý ảnh hưởng đến doanh thu khách. (ta có thể cần nhiều feature của tháng 4 để dự đoán cho tháng 8)

### Phần 2: Kiểm tra tương quan

Các đặc trưng cùng loại tương quan đáng kể với nhau.

Các đặc trưng về doanh thu có tương quan đáng kể đến biến mục tiêu, tiếp đến là đặc trưng về cuộc gọi và ít ảnh hưởng là đặc trưng lượng dữ liệu sử dụng.

Đặc trừng về lượng dữ liệu sử dụng đặc biệt ít tương quan đến các đặc trưng khác, có thể lưu lượng dữ liệu không ảnh hưởng đáng kể đến việc phán đoán.

## 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 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.

### Phần 3: Feature Engineering

Ta sẽ bổ sung thêm một số đặc trưng categorical

* decrease\_arpu\_123: True nếu doanh thu tháng 7 giảm, giảm được định nghĩa nếu doanh thu tháng này nhỏ hơn trung bình 3 tháng trước
* decrease\_data\_123: True nếu doanh thu tháng 7 giảm, giảm được định nghĩa nếu data tháng này nhỏ hơn trung bình 3 tháng trước
* decrease\_call\_123: tương tự

|  |
| --- |
|  |

Có sự tương quan, nhưng không đáng kể.

### Phần 4: Huấn luyện mô hình với Random forest và XGBOOST

|  |  |  |  |
| --- | --- | --- | --- |
|  | MAE | RMSE | R-square |
| Random forest | 55.78 | 121.57 | 0.488 |
| XGBOOST | 48.41 | 118.13 | 0.530 |

XGBoost cho kết quả tốt hơn hẳn, Randomforest bị biased không phán đoán đúng cho những trường hợp arpu ở mức thấp (129-200 nghìn). XGBOOST có phân phối dự đoán của arpu\_5 sát với thực tế hơn

|  |
| --- |
|  |
|  |

## Phần: Sử dụng kết quả cho các bài toán đề xuất

### Phần 1: Bài toán cảnh báo doanh thu giảm từ khách hàng có mức tiêu cao

* Ví dụ: decrease\_arpu\_123: True nếu doanh thu tháng 7 giảm, giảm được định nghĩa nếu doanh thu tháng này nhỏ hơn trung bình 3 tháng trước, ở đây là doanh thu tháng 7 nhỏ hơn trung bình doanh thu tháng 4,5,6.
* Vậy ta sẽ tính mean\_arpu\_234 và kiểm tra xem arpu\_5 (tháng 8) có nhỏ hơn trung bình doanh thu tháng 5,6,7 hay không
* Nếu có ta sẽ gắn nhãn thuê bao đó là giảm doanh thu và đưa ra cách tác động

Giảm được gán là positive, không giảm được gán là negative. Kết quả sử dụng cho bài toán này:

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Thực | |
|  |  | Giảm | Không giảm |
| Predicted | Giảm | 3808 | 1924 |
| Không giảm | 7891 | 10799 |

Precision: 66,4%

Recall: 32.5%

Accuracy: 59.81%

Nhận xét: Kết quả bài toán chưa cao, khó áp dụng được khi chi phí phán đoán sai cho việc không giảm là lớn, ta sẽ mất khách hàng về lâu dài

### Phần 2: Bài toán phân khúc khách hàng

Phân khúc các thuê bao thành từng nhóm với phân cụm Kmeans nhằm phục vụ những gói mới, ưu đãi mới phù hợp.

Ta phân nhóm với data ban đầu trước. Sau đó ta thực hiện phân nhóm với tập test từ bài toán ban đầu. Cuối cùng, so sánh thuê bao này có được phân nhóm đúng với nhóm ở data ban đầu hay không. Kết quả như sau:

* Lượng người được phân nhóm đúng: 24291/24409
* Độ chính xác: 99.52%

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Arpu\_1 | Arpu\_2 | Arpu\_3 | Arpu\_4 | Call\_1 | Call\_2 | Call\_3 | Call\_4 | Pred\_arpu\_5 |
| Cụm |  |  |  |  |  |  |  |  |  |
| 0 | 254.7 | 247.7 | 262.3 | 272.5 | 110.0 | 136.1 | 145.2 | 165.0 | 274.2 |
| 1 | 167.6 | 157.2 | 155.1 | 143.6 | 77.13 | 87.64 | 89.63 | 95.85 | 154.0 |
| 2 | 375.8 | 367.6 | 387.3 | 390.3 | 139.0 | 181.6 | 200.3 | 232.7 | 643.2 |
| 3 | 336.1 | 333.1 | 371.3 | 389.7 | 117.3 | 156.2 | 174.4 | 203.9 | 482.5 |
| 4 | 0 | 0 | 387.6 | 390.3 | 0 | 0 | 0 | 0 | 2501.1 |
| 5 | 205.2 | 194.2 | 201.4 | 197.8 | 96.8 | 117.4 | 124.6 | 138.4 | 206.3 |
| 6 | 318.3 | 320.6 | 337.7 | 356.9 | 114.2 | 150.1 | 163.5 | 185.9 | 367.4 |

Có nhóm 4 bất thường, có thể không cần quan tâm

Nhóm 1: doanh thu mức thấp, được dự đoán giảm ít, và lượng cuộc gọi ít. => tặng

Nhóm 2: doanh thu mức cao, được dự đoán tăng nhiều, và lượng cuộc gọi lớn.

Nhóm 3: doanh thu mức cao, được dự đoán tăng nhiều, lượng cuộc gọi lớn. Tuy nhiên kém hơn nhóm 2 1 ít.

* Ưu đãi cho các gói cước cao, gói đi kèm

Nhóm 0: doanh thu mức trung, lượng cuộc gọi cao và doanh thu dự đoán tăng ít.

Nhóm 5: doanh thu mức trung, lượng cuộc gọi ở mức trung, được dự đoán tăng ít.

Nhóm 6: doanh thu mức cao, lượng cuộc gọi ở mức trung, doanh thu được dự đoán tăng ít.