Tiền xử lý cho EDA

* Loại bỏ một số trường không cần thiết: 'min\_tuoi', 'idnumber', 'co\_goi', 'khong\_tuong\_tac', 'accalo', 'is\_onsong', 'BTS\_NAME', 'ngay\_sinh' (vì ta đã biết tuổi đi kèm), 'uutientap4', 'PTC', ‘num\_of\_id’.
* Chuẩn hóa lại giá trị của ‘gender’:

'male': 'male',

'Male': 'male',

'nam': 'male',

'Nam': 'male',

'NAM / M': 'male',

'female': 'female',

'Female': 'female',

'Nữ': 'female',

'nữ': 'female',

'nu': 'female',

'-': 'notKnown',

* Đưa về đơn vị nghìn đồng với các đặc trưng vì độ lớn của giá trị: 'NAP\_THE\_T','TKC\_T', 'TKC\_THOAI\_T', 'TKC\_THOAI\_ONNET\_T','TKC\_THOAI\_OFFNET\_T','TKC\_SMS\_T', 'TKC\_SMS\_ONNET\_T','TKC\_SMS\_OFFNET\_T','TKC\_DATA\_T','TKC\_KHAC\_T','TKC\_VAS\_T',

                  'NAP\_THE\_T\_1','TKC\_T\_1', 'TKC\_THOAI\_T\_1','TKC\_THOAI\_ONNET\_T\_1','TKC\_THOAI\_OFFNET\_T\_1','TKC\_SMS\_T\_1', 'TKC\_SMS\_ONNET\_T\_1','TKC\_SMS\_OFFNET\_T\_1','TKC\_DATA\_T\_1','TKC\_KHAC\_T\_1','TKC\_VAS\_T\_1',

                'NAP\_THE\_T\_2','TKC\_T\_2', 'TKC\_THOAI\_T\_2','TKC\_THOAI\_ONNET\_T\_2','TKC\_THOAI\_OFFNET\_T\_2','TKC\_SMS\_T\_2', 'TKC\_SMS\_ONNET\_T\_2','TKC\_SMS\_OFFNET\_T\_2','TKC\_DATA\_T\_2','TKC\_KHAC\_T\_2','TKC\_VAS\_T\_2',

'TKC\_MYTV\_T','TKC\_MYTV\_T\_1','TKC\_MYTV\_T\_2'

(TKKM giữ nguyên vì giá trị nhỏ nếu chia)

* Đưa về đơn vị TB với các đặc trưng vì độ lớn của giá trị data: 'LL\_DATA\_T','LL\_DATA\_T\_1','LL\_DATA\_T\_2'
* Bổ sung “notKnown” cho một số trường không ghi nhận thông tin chủ thuê bao: 'tuoi', 'gender', 'PROVINCE\_CODE', 'PBH', 'GEO\_CITY\_NAME', 'GEO\_CNTY\_NAME' (Chưa được dùng), 'APP\_NAME' (Không đăng ký gói), ‘simslot’ (Không ghi nhận thiết bị đi kèm)
* Bổ sung “notIncluded” cho các trường: 'device\_type', ‘'device\_spam' vì SIM không kích hoạt trên các thiết bị này
* Điền 0 nếu không ghi nhận thông tin cho các trường : 'amount\_nap\_the\_t6', 'voice\_ngoai\_mang\_t6', 'voice\_nm\_t6', 'voice\_di\_ngoaimang\_t6', 'voice\_den\_ngoaimang\_t6', 'voice\_di\_nm\_t6', 'voice\_den\_nm\_t6', 'sum\_voice\_t6', 'amount\_nap\_the\_t7', 'voice\_ngoai\_mang\_t7', 'voice\_nm\_t7', 'voice\_di\_ngoaimang\_t7', 'voice\_den\_ngoaimang\_t7', 'voice\_di\_nm\_t7', 'voice\_den\_nm\_t7', 'sum\_voice\_t7'.
* Điền 0 nếu thuê bao không đăng ký các dịch vụ: 'is\_sim\_bundle', ‘is\_fiber', 'is\_tieudung\_km', 'package\_km', 'package\_km\_t\_1'
* Ghi nhận “notSub” nếu thuê bao không đăng ký gói ở các trường: 'SERVICE\_CODE', 'SERVICE\_CODE\_MUAGOI', 'package\_homecombo'
* Điền 0 với “P1\_SONGAY” nếu không đăng ký gói ngày trong 'SERVICE\_CODE'. Điền 0 với “P2\_CHUKY” nếu không đăng ký gói trong 'SERVICE\_CODE\_MUAGOI', và 0,1 với những gói theo ngày ghi nhận trong 'SERVICE\_CODE\_MUAGOI'.

Phân tích tổng quan

Trên các trường TKC: hầu hết ghi nhận phân phối lệch trái, ngoại lai xuất hiện nhiều.

+ Thoại ngoại mạng: 50% mẫu từ 0 đến 300 đồng, 75% là 6 nghìn. Tăng nhẹ từ t7-t9

+ Thoại nội mạng: 50% mẫu là 0 đồng, 75% là 2,6 nghìn. Tăng nhẹ từ t7-t9

+ SMS nội mạng: 75% là 0 đồng.

+ SMS ngoại mạng: 75% là 0 đồng.

+ Data: 50% là từ 0 đến 400 đồng, 75% là 8 nghìn, giá trị này giảm dần từ t7-t9.

+ Khác: 75% là 0 đồng. Không đổi qua các tháng

+ VAS: 50% là 0 đồng, điểm tứ phân vị thứ 3 là 1 nghìn. Không đổi qua các tháng

Trên các trường TKKM: Giá trị thấp, 75% giá trị ghi nhận đều bằng 0. Ngoại lai xuất hiện nhiều.

Trên các trường LL: Phân phối tương tự như ở các trường TKC.

+ LL data sử dụng 50% thuê bao là từ 0 đến 2,45GB. 75% là 3,54GB. Giảm từ t7-t9

+ LL thoại nội mạng đến: 50% là ~55 phút. Không biến động nhiều qua các tháng

+ LL thoại nội mạng đi: 50% là rất ít 0 phút, giảm dần qua các tháng.

+ LL thoại ngoại mạng đi: 50% là 12 phút, 75% là 261 phút. giảm dần qua các tháng

+ SMS nội mạng: 75% là 0. Hông thay đổi qua các tháng

+ SMS ngoại mạng: 50% là 0, 75% là 3 SMS. Hông đổi qua các tháng

Tuy nhiên, có sự khác biệt lớn ở đặc trưng về 'LL\_SMS\_T', 'LL\_SMS\_T\_1', 'LL\_SMS\_T\_2' khi có xuất hiện phần đồi của phân phối.

A graph with blue lines

Description automatically generated

Ta sẽ phân nhóm thuê bao theo đặc trưng này để tìm ra những điểm khác biệt.

Về tương quan (theo từng tháng)

* Với T-1: NAP\_THE và TKC; TKKM và TKKM\_THOAI. TKC\_THOAI\_OFFNET và TKC\_THOAI. TKC SMS và TKC\_SMS\_ONNET. LL\_OFFNET\_OG với TKC\_THOAI\_OFFNET, LL\_SMS\_ONNET với TKC\_SMS và TKC\_SMS\_ONNET
* Với T-2: NAP\_THE và TKC; TKKM và TKKM\_THOAI. TKC SMS và TKC\_SMS\_ONNET, LL\_OFFNET\_OG với TKC\_THOAI\_OFFNET, LL\_SMS\_ONNET với TKC\_SMS và TKC\_SMS\_ONNET, LL\_SMS với TKC\_SMS và TKC\_SMS\_ONNET.
* Với T:

Về tương quan với nhóm đặc trưng voice đến, đi nội ngoại mạng.

* Có tương quan lớn giữa các đặc trưng cùng tên theo từng tháng (thay bằng rung bình), và voice đến nội mạng với tổng voice nội mạng, tổng voice ngoại mạng với voice đến và đi ngoại mạng. Tổng voice tương quan cao nhất với tổng voice ngoại mạng.
* Theo thời gian, tương quan lớn nằm ở lưu lượng. Ngoài việc tương quan theo thời gian giữa các đặc trưng cùng tên, thì LL\_THOAI\_T2 tương quan mạnh với LL\_ONNET\_IC các tháng và ngược lại. Điều này sẽ giúp ta gợi ý về tỉ lệ chuyển đổi của 2 đặc trưng này. LL\_SMS ONNET là tương quan kém nhất.