Mining frequent Itemsets and Association Rules

Họ và tên: Ngô Đăng Khoa

Môn học: Khai thác dữ liệu và ứng dụng - 19KHDL

Lecturer: Lê Ngọc Thành - **TA:** Nguyễn Thái Vũ

I. MÔ TẢ VÀTIỀN XỬ LÝ DỮ LIỆU

1. Khám phá dữ liệu

	State	Account Length	Area Code	Phone	Int'l Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls	Churn?
0	KS	128	415	382-4657	no	yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10.0	3	2.70	1	False.
1	OH	107	415	371-7191	no	yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.70	1	False.
2	NJ	137	415	358-1921	no	no	0	243.4	114	41.38	121.2	110	10.30	162.6	104	7.32	12.2	5	3.29	0	False.
3	OH	84	408	375-9999	yes	no	0	299.4	71	50.90	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False.
4	OK	75	415	330-6626	yes	no	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False.

- Dữ liệu gồm có 333 dòng và 21 cột
- Ý nghĩa của mỗi cột được mô tả như sau:
    - **State**: categorical, Tên viết tắt của 50 bang và thủ đô Washington D.C
    - **Account length:** integer-valued, Thời gian sử dụng kể từ ngày kích hoạt dịch vụ của khách hàng
    - **Area code**: categorical, Mã vùng
    - **Phone**: Số điện thoại, đại diện cho ID của mỗi khách hàng.
    - **Int'l Plan**: dichotomous categorical, Phân loại khách hàng có sử dụng dịch vụ quốc tế hay không
    - **VMail Plan**: dichotomous categorical, Phân loại khách hàng có sử dụng Voice mail hay không.
    - **VMail Message**: integer-valued, Số lượng voice mail của khách hàng.
    - **Day Mins**: Số phút khách hàng sử dụng vào ban ngày
    - **Day Calls**: Số cuộc gọi của khách hàng vào ban ngày
    - **Day Charge**: Cước phí gọi vào ban ngày, có thể phụ thuộc vào 2 biến ở trên
    - **Eve Mins**: Số phút khách hàng sử dụng vào chiều tối
    - **Eve Calls**: Số cuộc gọi của khách hàng vào chiều tối
    - **Eve Charge**: Cước phí gọi vào chiều tối, có thể phụ thuộc vào 2 biến ở trên
    - **Night Mins**: Số phút khách hàng sử dụng vào ban đêm
    - **Night Calls**: Số cuộc gọi của khách hàng vào ban đêm
    - **Night Charge**: Cước phí gọi vào ban đêm, có thể phụ thuộc vào 2 biến ở trên
    - **Intl Mins:** Số phút gọi quốc tế
    - **Intl Calls**: Số cuộc gọi quốc tế của khách hàng
    - **Intl Charge**: Cước phí gọi quốc tế, có thể phụ thuộc vào 2 biến ở trên
    - **CustServ Calls**: Số cuộc gọi đến tổng đài dịch vụ khách hàng
    - **Churn?**: Phân loại xem khách hàng đã hủy dịch vụ hay chưa
- Các cột dữ liệu dạng numeric:

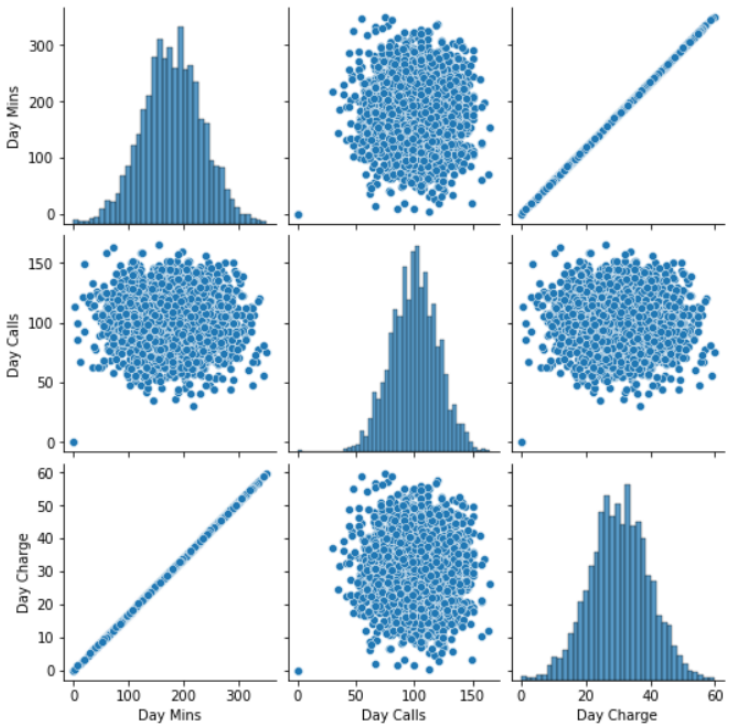
	Account Length	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls
missing_ratio	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.00	0.0	0.0	0.00	0.0	0.0	0.0	0.0
min	1.0	0.0	0.0	0.0	0.00	0.0	0.0	0.00	23.2	33.0	1.04	0.0	0.0	0.0	0.0
max	243.0	51.0	350.8	165.0	59.64	363.7	170.0	30.91	395.0	175.0	17.77	20.0	20.0	5.4	9.0

- Các cột dữ liệu dạng Categorical:

	State	Area Code	Phone	Int'l Plan	VMail Plan	Churn?
missing_ratio	0.0	0.0	0.0	0.0	0.0	0.0
num_diff_vals	51	3	3333	2	2	2
diff_vals	[AK, AL, AR, AZ, CA, CO, CT, DC, DE, FL, GA, H...	[408, 415, 510]	[327-1058, 327-1319, 327-3053, 327-3587, 327-3...	[no, yes]	[no, yes]	[False., True.]

2. Tiền xử lý dữ liệu

- Vấn đề cần tiền xử lý:
  - Chúng ta cần phải tránh đưa các biến tương quan với nhau vào trong mô hình data mining vì khi sử dụng những biến tương quan sẽ overemphasize các thành phần dữ liệu làm cho mô hình trở nên không ổn định và đưa ra những kết quả không tốt.
- Qua phần khám phá dữ liệu, ta có cảm giác như các biến **minutes, calls, charge** có vẻ như tương quan với nhau (có thể **charge** là một hàm dựa trên hai biến còn lại) nên đầu ta sẽ quan sát chúng trước qua đồ thị quan hệ dưới đây:



- Qua biểu đồ trên ta thấy được các mối quan hệ giữa 3 biến trên. Đầu tiên ta thấy được rằng không có mỗi quan hệ nào giữa biến **Day Calls** với **Day Mins** và **Day Charge**. Nhưng thay vào đó thấy được mối tương quan tuyến tính giữa 2 biến **Day Mins** với **Day Charge** khá rõ ràng nên chúng ta sẽ loại bỏ biến (tùy ý) **Day Charge**. Tương tự với các trường hợp còn lại ta cũng sẽ loại bỏ các cột **Eve Charge, Night Charge, Intl Charge.**
- Trong file **churn_description.pdf** có mô tả một sự bất thường trong trường dữ liệu **Area Code**, trường này chứa 3 giá trị riêng biệt 408, 415, 510 - Đây đều là những mã vùng của bang California:

```
df['Area Code'].value_counts()
```
✓ 0.1s

```
415    1655
510     840
408     838
Name: Area Code, dtype: int64
```

- Điều này không có gì bất thường nếu như tất cả khách hàng đều sống ở Cali. Tuy nhiên trong bảng dữ liệu lại phân bố 3 mã vùng này ở tất cả các bang khác. Do đó chúng ta cần phải cảnh giác với trường dữ liệu này nên có lẽ sẽ không đưa trường dữ liệu này vào các giai đoạn tiếp theo

	State	415	510	408
0	WA	26	17	23
1	ND	28	15	19
2	NJ	34	19	15
3	AZ	36	13	15
4	WV	52	34	20
5	DC	27	13	14
6	KS	37	21	12
7	MD	39	15	16
8	TN	30	12	11
9	GA	21	18	15

- Ta có thể bỏ luôn các cột Categorical định danh của khách hàng như **Phone**, **State** để có thể tập trung vào việc khai thác mẫu phổ biến và luật kết hợp.

- Tiếp theo ta sẽ số hóa các cột categorical có giá trị Yes / No thành 0 và 1 để có thể dễ làm việc hơn.
- Cuối cùng ta sẽ quy định các cột numeric thành các dạng Binominal để có thể để đưa về transaction table theo dạng **Binarization**. Trong đó ta gom ta sẽ encode các cột **Calls** theo cột **Mins** bằng cách nếu cột **Min** có giá trị khác 0 thì cột **Calls** sẽ có giá trị là 1 tương ứng với việc có sử dụng dịch vụ gọi đó. Tương tự như với các Service còn lại.

	Int'l Plan	VMail Plan	VMail Message	Day Calls	Eve Calls	Night Calls	Intl Calls	CustServ Calls	Churn?
0	0	1	1	1	1	1	1	1	0
1	0	1	1	1	1	1	1	1	0
2	0	0	0	1	1	1	1	0	0
3	1	0	0	1	1	1	1	1	0
4	1	0	0	1	1	1	1	1	0
...	...	...	...	...	...	...	...	...	...
3328	0	1	1	1	1	1	1	1	0
3329	0	0	0	1	1	1	1	1	0
3330	0	0	0	1	1	1	1	1	0
3331	1	0	0	1	1	1	1	1	0
3332	0	1	1	1	1	1	1	0	0

- Quan sát tần suất của bảng trên, Có vẻ như khi đưa về dạng **Binarization,** ta thấy được thêm mối quan hệ giữa 2 cột **Vmail Plan** và **Vmail Message** đó là chỉ có những người có đăng ký **VMail Plan** thì mới có thể sử dụng **Vmail Message** nên ta sẽ loại bỏ đi một trong 2 cột này.

```
Int'l Plan        323
VMail Plan        922
VMail Message     922
Day Calls        3331
Eve Calls        3331
Night Calls      3333
Intl Calls       3315
CustServ Calls   2636
Churn?            483
dtype: int64
```

```
df[df['VMail Plan'] != df['VMail Message']]
```
✓ 0.1s

	Int'l Plan	VMail Plan	VMail Message	Day Calls	Eve Calls	Night Calls	Intl Calls	CustServ Calls	Churn?

Không có trường hợp nào thể hiện ngược lại điều trên

- Bảng dữ liệu sau khi tiền xử lý dữ liệu:

	Int'l Plan	VMail Plan	Day Calls	Eve Calls	Night Calls	Intl Calls	CustServ Calls	Churn?
0	0	1	1	1	1	1	1	0
1	0	1	1	1	1	1	1	0
2	0	0	1	1	1	1	0	0
3	1	0	1	1	1	1	1	0
4	1	0	1	1	1	1	1	0
...	...	...	...	...	...	...	...	...
3328	0	1	1	1	1	1	1	0
3329	0	0	1	1	1	1	1	0
3330	0	0	1	1	1	1	1	0
3331	1	0	1	1	1	1	1	0
3332	0	1	1	1	1	1	0	0

II. KHAI THÁC TẬP PHỔ BIẾN VÀ LUẬT KẾT HỢP

1. Thuật toán

- Để có thể khai thác mẫu phổ biến và luật kết hợp, em đã sử dụng thuật toán **Apriori** và **ASSOCIATION RULES** đã được cài đặt trong file **apriori.py** và import vào trong Jupyter Notebook. Đối với luật kết hợp, em có sử dụng thêm độ đo lift để đánh giá mối quan hệ giữa các itemset với nhau từ đó có thể sinh ra những luật tin cậy hơn

2. Câu hỏi và phân tích

- **Câu hỏi 1: Khách hàng thường sử dụng chung những dịch vụ nào? Để từ đó có thể đưa ra những chính sách ưu đãi hợp lý**
  - Ở câu hỏi này em sẽ đặt minsup là 9% ứng với khoảng 300 khách hàng để có thể quan sát mẫu của tất cả các dịch vụ (dịch vụ có tần suất thấp nhất là **Int'l Plan** 323 người sử dụng ~ 9,6%). Tuy minsup thấp nhưng với số lượng khách hàng như vậy cũng khá nhiều để quan tâm. Còn thông số minconf = 80% cùng với độ đo lift > 1 để tăng độ tin cậy.

```
108  ['VMail Plan'] ---> ['Day Calls', 'Eve Calls', 'Intl Calls'], conf = 0.996, lift = 1.002, sup = 0.275427542754274
```

  - Ví dụ như ta có tập luật giữa **'Eve Calls'**, **'Day Calls'**, **'Vmail Plan'** và **'Intl Calls'** có độ trợ khá cao (~ 27,5%) cũng như các thông số về conf và lift đều tốt nên em nghĩ cần có thêm những ưu đãi chú trọng vào 3 dịch vụ trên. Ví dụ như nếu như có đăng ký dịch vụ Voice Mail thì cần giảm giá cước cho các cuộc gọi kia.

- Tương tự với cả gói dịch vụ quốc tế **Int'l Plan**

```
96    ["Int'l Plan"] ---> ['Day Calls', 'Eve Calls', 'Intl Calls'], conf = 1.0, lift = 1.006, sup = 0.0969096909690972
```

- Ngoài ra em cũng muốn biết thêm là tại sao độ trợ của gói cước gọi quốc tế **Int'l Plan** không nhiều nhưng lại có số khách hàng gọi quốc tế (**Intl Calls)** khá lơn. Liệu có vấn đề gì nằm ở gói cước dịch vụ quốc tế hay không

- **Câu 2: Có dịch vụ nào dẫn đến việc khách hàng từ bỏ công ty (Churn) hay không?**

  - Qua tập luật kết hợp ở trên, tuy Churn có xuất hiện ở một số luật kết hợp nhưng lại không có luật nào suy ra là khách hàng sẽ bỏ công ty vì một dịch vụ nào hết.

- **Câu 3: Liệu có những dịch vụ nào thường xuyên gọi đến trung tâm chăm sóc khách hàng không? Từ đó công ty sẽ chú ý hơn đến những thắc mắc của khách hàng**.
  - Qua luật kết hợp trên đa số khách hàng sẽ gọi đến trung tâm hỗ trợ đối với những dịch vụ gọi **Days, Evening, Night** và **Churn**. Từ đó công ty cần phải xem xét tìm hiểu kĩ đang có những vấn đề đang tồn đọng ở những dịch vụ trên.

```
178   ['Churn?', 'Day Calls', 'Eve Calls', 'Intl Calls', 'Night Calls'] ---> ['CustServ Calls'], conf = 0.809, lift = 1.023, sup = 0.117011701
```

III. TÀI LIỆU THAM KHẢO

- Slide lý thuyết
- Textbook: J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition.
- File churn_description.pdf
- https://www.techtarget.com/searchbusinessanalytics/definition/association-rules-in-data-mining