TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP. HCM KHOA CÔNG NGHỆ THÔNG TIN

DATA MINING

BÁO CÁO LAB2

SINH VIÊN THỰC HIỆN:

1712919 Lê Văn Vũ

1712502 Trần Quang Huy

GV LÝ THUYẾT/ HD THỰC HÀNH:

Thầy Lê Hoài Bắc

Thầy Nguyễn Ngọc Đức

Thầy **Dương Nguyễn Thái Bảo**

Thầy **Hoàng Xuân Trường**

MỤC LỤC

I. Data:

Tập dữ liệu churn.txt có 3333 mẫu, 21 thuộc tính:

- State: Các bang thuộc Columbia. Thuộc tính định danh.
- Account length: Thời gian tính từ kích hoạt tài khoản. Số nguyên
- Area Code: mã khu vực. Thuộc tính định danh.
- Phone: Số điện thoại của khách hàng. Dùng để làm ID cho khách hành.
- Vmail Plan: Có sử dụng dịch vụ Vmail hay không? Thuộc tính nhị phân.
- Intl Plan: Có sử dụng dịch vụ quốc tế hay không? Thuộc tính nhị phân.
- Vmail Message: Số tin nhắn. Số nguyên.
- Day Mins: Số phút gọi trong ngày. Số nguyên
- Day Calls: Số cuộc gọi trong ngày. Số nguyên.
- Day Charge: Phí gọi trong ngày. Số nguyên.
- Eve Mins: Số phút gọi vào buổi chiều. Số nguyên
- Eve Calls: Số cuộc gọi vào buổi chiều. Số nguyên.
- Eve Charge: Phí gọi vào buổi chiều. Số nguyên.
- Night Mins: Số phút gọi vào ban đêm. Số nguyên
- Night Calls: Số cuộc gọi vào ban đêm. Số nguyên.
- Night Charge: Phí gọi vào ban đêm. Số nguyên.
- Intl Mins: Số cuộc phút gọi quốc tế. Số nguyên.
- Intl Calls: Số cuộc gọi quốc tế. Số nguyên.
- Intl Charge: Phí gọi quốc tế. Số nguyên,
- CustServ Calls: Cuộc gọi đến dịch vụ chăm sóc khách hàng. Số nguyên.

(kết quả khi dùng dtype:

```
Unnamed: 0          int64
State              object
Account Length      int64
Area Code           int64
Phone              object
Int'l Plan         object
VMail Plan         object
VMail Message       int64
Day Mins          float64
Day Calls           int64
Day Charge        float64
Eve Mins          float64
Eve Calls           int64
Eve Charge        float64
Night Mins        float64
Night Calls         int64
Night Charge      float64
Intl Mins         float64
Intl Calls          int64
Intl Charge       float64
CustServ Calls      int64
Churn?             object
dtype: object
```

➔Thuộc tính lớp là : **Churn?**.là thuộc tính nhị phân: cho biết khách hàng đó có ngưng sử dụng dịch vụ hay không?

Mô tả dữ liệu:

	Unnamed: 0	Account Length	Area Code	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
count	3333.00000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000
mean	1666.00000	101.064806	437.182418	8.099010	179.775098	100.435644	30.562307	200.980348	100.114311	17.083540	200.872037	100.107711	9.039325	10.237294	4.479448	2.764581	1.562856
std	962.29855	39.822106	42.371290	13.688365	54.467389	20.069084	9.259435	50.713844	19.922625	4.310668	50.573847	19.568609	2.275873	2.791840	2.461214	0.753773	1.315491
min	0.00000	1.000000	408.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	23.200000	33.000000	1.040000	0.000000	0.000000	0.000000	0.000000
25%	833.00000	74.000000	408.000000	0.000000	143.700000	87.000000	24.430000	166.600000	87.000000	14.160000	167.000000	87.000000	7.520000	8.500000	3.000000	2.300000	1.000000
50%	1666.00000	101.000000	415.000000	0.000000	179.400000	101.000000	30.500000	201.400000	100.000000	17.120000	201.200000	100.000000	9.050000	10.300000	4.000000	2.780000	1.000000
75%	2499.00000	127.000000	510.000000	20.000000	216.400000	114.000000	36.790000	235.300000	114.000000	20.000000	235.300000	113.000000	10.590000	12.100000	6.000000	3.270000	2.000000
max	3332.00000	243.000000	510.000000	51.000000	350.800000	165.000000	59.640000	363.700000	170.000000	30.910000	395.000000	175.000000	17.770000	20.000000	20.000000	5.400000	9.000000

II. Phân tích dữ liệu (EDA) và các tác cụ tiền xử lí.

1. Phân tích dữliệu

a. Mức độ tương quan giữa các loại các thuộc tính

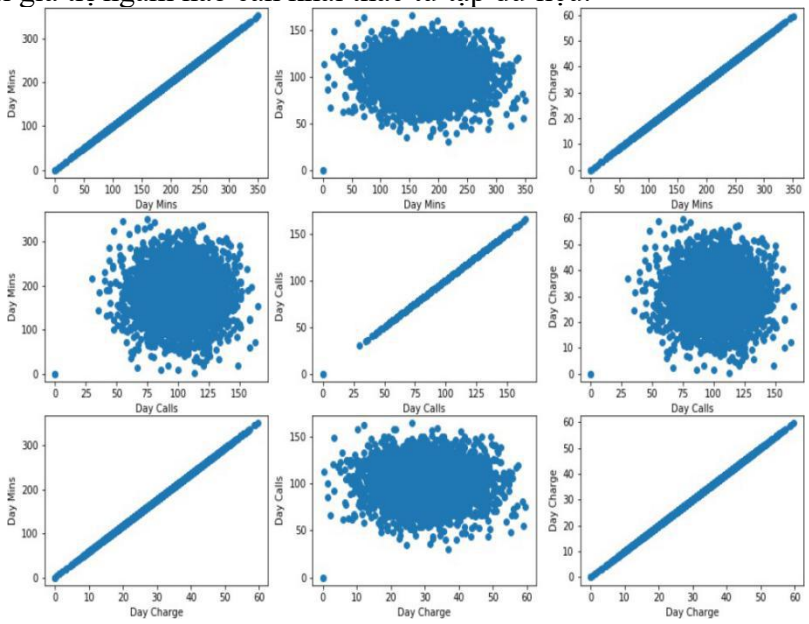
Giá trị từ nằm trong khoảng 0 -> 1 thể hiện 2 thuộc tính có mối tương quan tăng dần. Ngược lại trong khoảng -1 <- 0 thể hiện 2 thuộc tính không có mối tương quan tăng dần.

	Account Length	VMail Message	Day Mins	Day Calls	Eve Mins	Eve Calls	Night Mins	Night Calls	Intl Mins	Intl Calls	CustServ Calls
Account Length	1.000000	0.021560	-0.025046	-0.001039	-0.028508	0.031134	-0.025191	-0.038920	0.027341	0.036169	0.036482
VMail Message	0.021560	1.000000	-0.000062	-0.000077	0.026548	-0.005431	-0.042074	0.007942	0.010150	0.054295	-0.035468
Day Mins	-0.025046	-0.000062	1.000000	0.002726	0.010370	0.043009	0.000500	0.006747	0.006007	-0.003297	-0.062942
Day Calls	-0.001039	-0.000077	0.002726	1.000000	0.037154	0.033831	-0.012137	-0.048819	0.029772	0.021031	-0.005004
Eve Mins	-0.028508	0.026548	0.010370	0.037154	1.000000	0.033168	0.018090	0.019662	-0.043247	-0.020019	-0.007080
Eve Calls	0.031134	-0.005431	0.043009	0.033831	0.033168	1.000000	0.005948	0.051961	0.037169	0.038633	0.007289
Night Mins	-0.025191	-0.042074	0.000500	-0.012137	0.018090	0.005948	1.000000	-0.050477	-0.044155	0.004455	-0.031653
Night Calls	-0.038920	0.007942	0.006747	-0.048819	0.019662	0.051961	-0.050477	1.000000	-0.018664	-0.005541	0.008355
Intl Mins	0.027341	0.010150	0.006007	0.029772	-0.043247	0.037169	-0.044155	-0.018664	1.000000	0.055955	-0.018565
Intl Calls	0.036169	0.054295	-0.003297	0.021031	-0.020019	0.038633	0.004455	-0.005541	0.055955	1.000000	0.000958
CustServ Calls	0.036482	-0.035468	-0.062942	-0.005004	-0.007080	0.007289	-0.031653	0.008355	-0.018565	0.000958	1.000000

Mức độ tương quan giữa các thuộc tính

*Vd: T*ương quan giữa Day Mins, Day Calls, Day charge

Nhìn vào biểu đồ trên, ta thấy được mối liên hệ giữa Day Mins và DayCharge là tuyến tính. Số cuộc gọi trong ngày càng nhiều thì số tiền phải chi trả càng nhiều → Việc này là hiển nhiên và không mang lại giá trị ngầm nào cần khai thác từ tập dữ liệu.

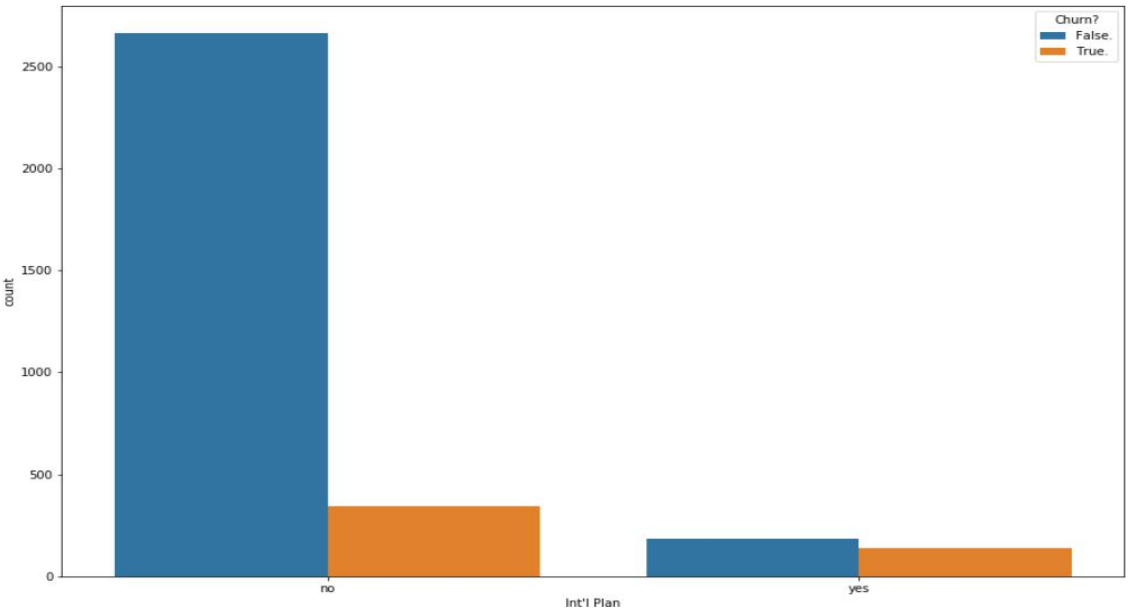


Biểu đồ thể hiện mối quan hệ giữa Day Mins, Day Calls và Day Charge

→ Chúng ta loại bỏ những trường charge trong tập dữ liệu.

b. Mối liên hệ của những thuộc tính rời rạc với thuộc tính phân lớp (churn):

❖ **International Plan:**

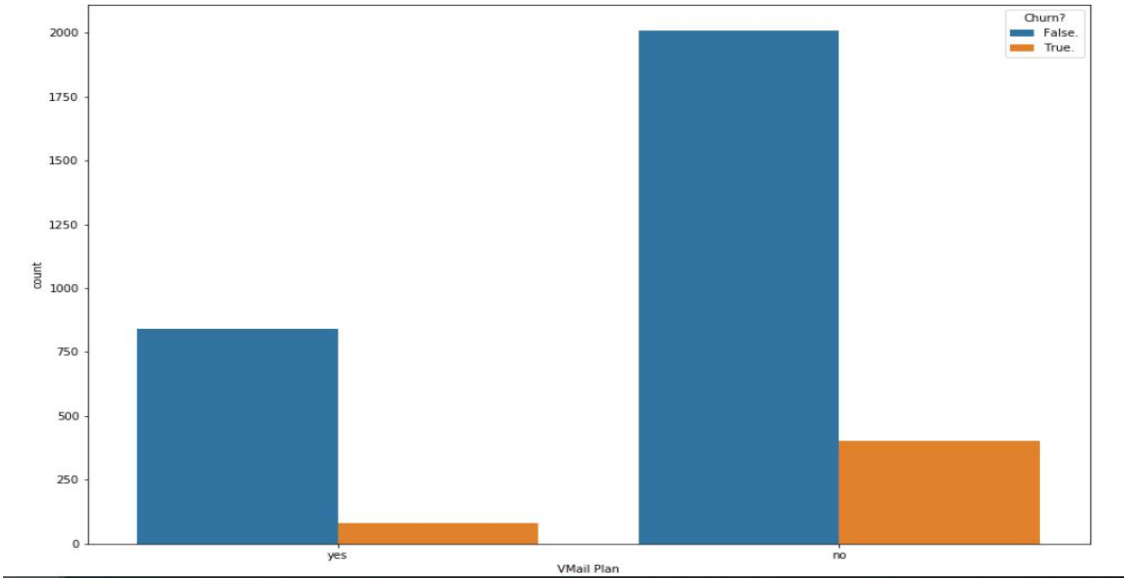


Quan hệ giữa Int'l Plan với Churn?

Biểu đồ cho ta thấy đa phần những khách hàng có sử dụng International Plan có xu hướng rời khỏi dịch vụ của công ty cao hơn. Theo thống kê, Có 323 người dùng sử dụng International Length này thì có tới 137 người sẽ rời khỏi dịch vụ, chiếm 42,4%. Trong khi những khách hàng không sử dụng International Length thì tỉ lệ này chỉ là 14,5%.

Từ đó ta chú tập trung vào nhóm khách hàng có sử dụng International Plan để đưa ra những ưu đãi, khuyến mãi để giữ chân những khách hàng này.

❖ **VoiceMail Plan:**

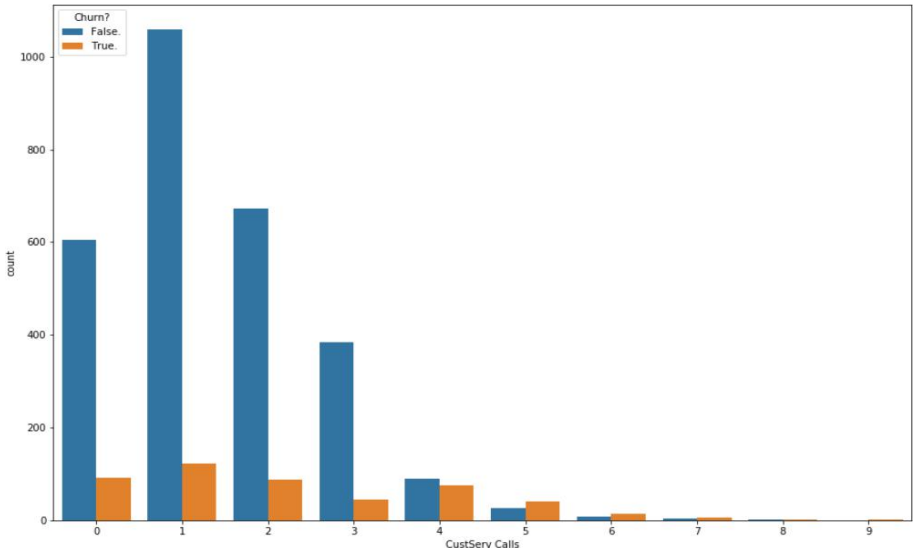


Biểu đồ để hiện quan hệ giữa Vmail Plan với Churn?

Biểu đồ trên cho ta thấy nhưng tỉ lệ khách hàng sử dụng VoiceMail Plan có rời khỏi dịch vụ thấp hơn so với tỉ lệ khách hàng không sử dụng VoiceMail Plan. Cụ thể 16,7% tỉ lệ khách hàng không sử dụng VoiceMail Plan là churn so với 8.7% của những khách hàng có sử dụng VoiceMail Plan.

Điều này chỉ ra rằng nếu chúng ta nâng cao, làm cho khách hàng dễ dàng tiếp cận với dịch vụ VoiceMail Plan thì có khả năng nâng cao lượng khách hàng trung thành.

❖ **Number of calls to customer service (CustServ Calls)**



Biểu đồ thể hiện mối quan hệ giữa CustServ Calls với Churn

Biểu đồ cho chúng ta thấy là tỉ lệ khách hàng mà thực hiện số cuộc gọi tới dịch vụ khách hàng từ 4 lần trở lên thường là những người sẽ rời bỏ dịch vụ. Mục tiêu là chúng ta theo dõi được số cuộc tới của mỗi khách hàng (có số cuộc gọi tới dưới 3 lần) để có thể cung cấp những đãi ngộ thích hợp để giữ chân những khách hàng này tiếp tục sử dụng dịch vụ.

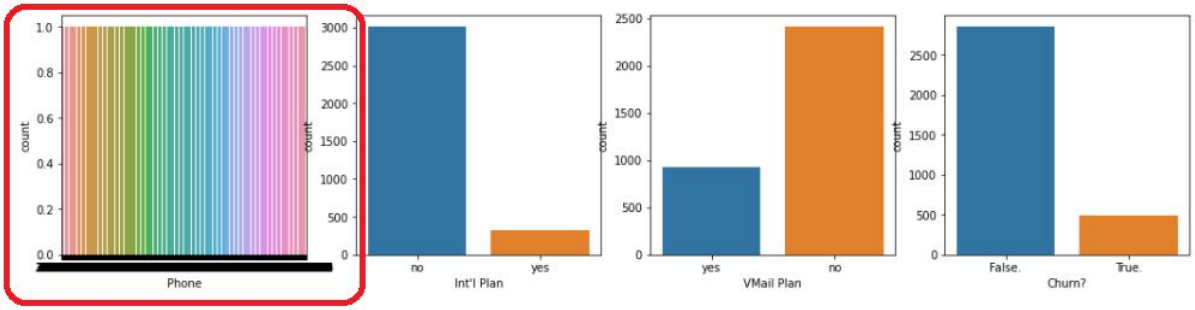
2. Các tác cụ tiền xử lí:

a. Loại bỏ các thuộc tính không mang giá trị sinh luật.

Các thuộc tính không mang lại giá trị sinh luật:

- **State:** Là trường dị thường. Không hiểu được ý nghĩa.
- **Account Length:** không có mối mối quan hệ rõ ràng với churn.
- **Area Code:** Là trường dị thường chỉ mang 3 giá trị 408, 415, 510.
- **Phone:** Không mang ý nghĩa.
- Như đã đề cặp ở trên các trường Charge không mang ý nghĩa nên ta loại đi, các trường này bao gồm:
  - **Day Charge**
  - **Eve Charge**
  - **Night Charge**
  - **Intl Charge**

Ví dụ: Thuộc tính Phone khi cho quá nhiều giá trị không có ý nghĩa so vớicac thuộc tính khác: (hình dưới)

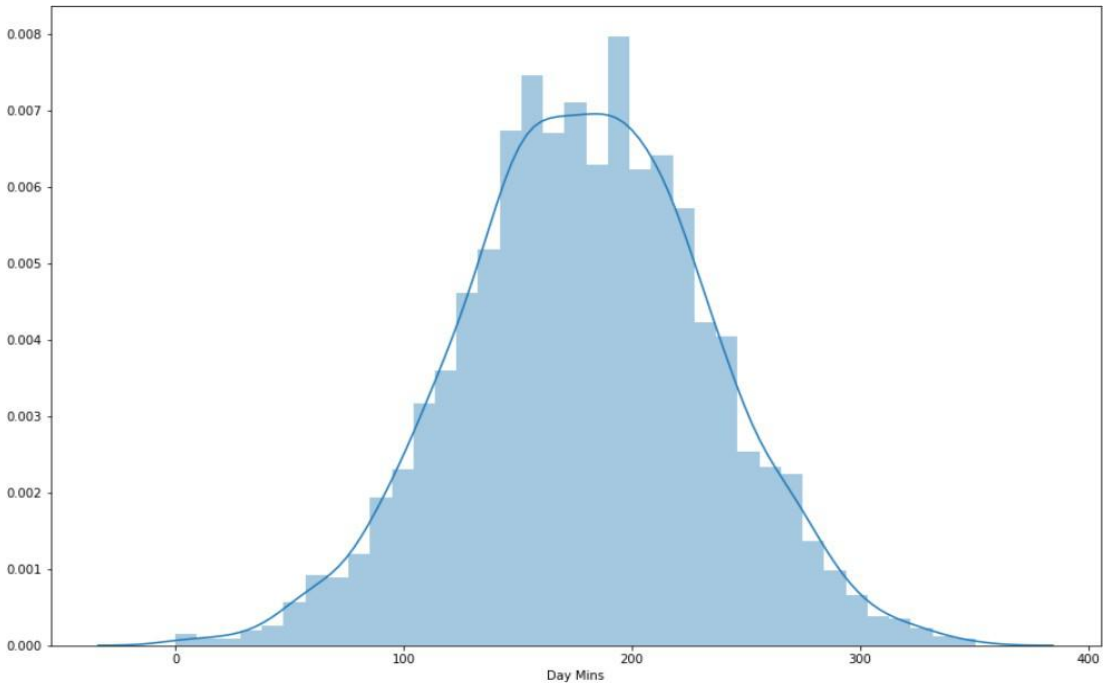


b. Phân lớp các thuộc tính liên tục:

Dựa vào mục *Mô tả dữ liệu: phần* ***I. Data***

- Chia các thuộc tính Day Mins, Day Calls, Eve Mins, Eve Calls, Night Mins, Night Calls thành 3 phần là low, median và high.

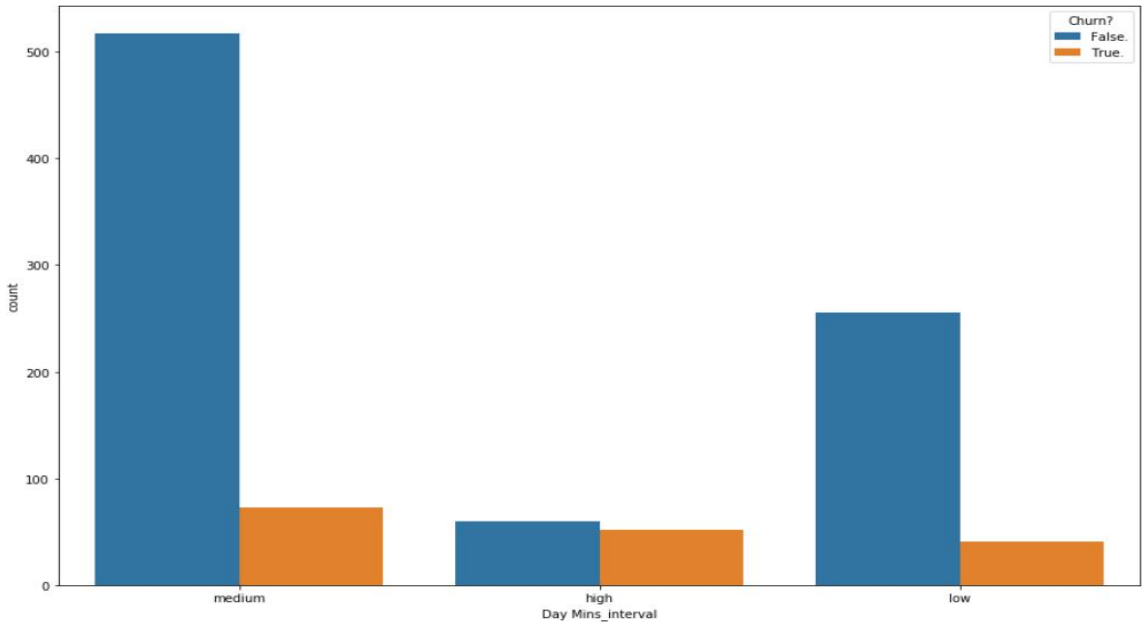
VD: với cách phân bố của thuộc tính Day Mins như sau, cột y là tần số xuất hiện, cột x là giá trị số phút gọi. Ta chia làm 3 khoảng, low=[0, 150), median=[150, 250), high=[250, max].



Hình 7: Tương quan giữa Day Mins với phân phối chuẩn.

- Thuộc tính Customer service calls (như *hình 7*) thành 2 khoảng giá trị low, high. Low=[0,3], high=[4,max].

- Thuộc tính Day Mins:



Quan hệ giữa Day Mins với Churn?

⇨ Từ biểu đồ ta thấy, với số phút gọi trong ngày cao thì tỉ lệ cao khách hàng này sẽ rời bỏ dịch vụ.

➔ Ta chỉ xét trường hợp số phút gọi trong ngày cao.

III. Code:

- File **pprocess.py** nhận đầu vào là file churn.txt, đầu ra là file: **ChurnProcessed.csv** chỉ chứa các thuộc tính và mẫu cần thiết để đưa vào Weka tìm ra các tập luật sinh.
- Chạy bằng tham số dòng lệnh: **pprocess.py churn.txt**

IV. Experiments

1. Mục đích phân tích dữ liệu:

Từ những phân tích dữ liệu trên, những luật kết hợp có ý nghĩa mà chúng ta muốn sinh ra (một số kết quả có tính chất bắc cầu dẫn đến người đó có khả năng là churner):

- Những luật dẫn đến người đó thực hiện cuộc gọi đến dịch vụ khách hàng.
- Những luật dẫn đến người đó sử dụng dịch vụ International plan.
- Những luật dẫn đến người đó sử dụng dịch vụ VoiceMail Plan (sử dụng ở độ tin cậy thấp)
- Những luật dẫn đến người đó có số phút gọi trong ngày cao.
- Những luật dẫn trực tiếp đến 1 người có khả năng là churn. *(rất khó trong tập dataset bị mất cân bằng)*

2. Thực nghiệm

Các phương pháp tiền xử lí

- Xóa các thuộc tính không cần thiết: State, Accound Length, VMail Message, Day Charge, Eve Charge, Night Charge, Intl Charge. Các chứng minh đã làm ở phần phân tích dữ liệu.

- Phân lớp cho các thuộc tính: Day Mins, Day Calls, Eve Mins, Eve Calls, Nignt Mins, Nignt Calls. Đã trình bày ở phần gom nhóm dữ liệu trong các thuộc tính thành các

nhóm.

➔ Được thực hiện trong file **pprocess.py**

	A	B	C	D	E	F	G	H	I	J	K
		Intl Plan	VMail Pla	Day Mins	Day Calls	Eve Mins	Eve Calls	Night Min	Night Call	CustServ	Churn?
2	0	no	yes	high	medium	medium	medium	medium	medium	low	False.
3	3	yes	no	high	low	low	medium	medium	medium	low	False.
4	9	yes	yes	high	medium	medium	medium	high	medium	low	False.
5	15	no	no	high	low	high	medium	medium	high	high	True.
6	61	no	yes	high	medium	medium	high	medium	medium	low	False.
7	66	yes	no	high	medium	medium	medium	medium	low	low	False.
8	76	no	no	high	medium	medium	medium	high	low	low	True.
9	93	no	no	high	medium	medium	medium	high	medium	low	False.
10	95	no	no	high	medium	low	medium	medium	high	low	False.
11	99	no	no	high	low	medium	high	high	medium	low	True.
12	106	no	yes	high	medium	medium	low	medium	medium	low	False.
13	117	no	no	high	medium	medium	low	medium	medium	low	True.
14	147	yes	no	high	medium	medium	high	low	medium	low	False.
15	149	no	yes	high	medium	medium	medium	high	medium	low	False.
16	154	no	no	high	medium	low	low	high	medium	high	False.
17	156	no	no	high	medium	medium	medium	medium	medium	low	True.
18	175	no	no	high	medium	medium	high	medium	medium	low	False.
19	184	yes	no	high	low	medium	high	low	high	low	False.

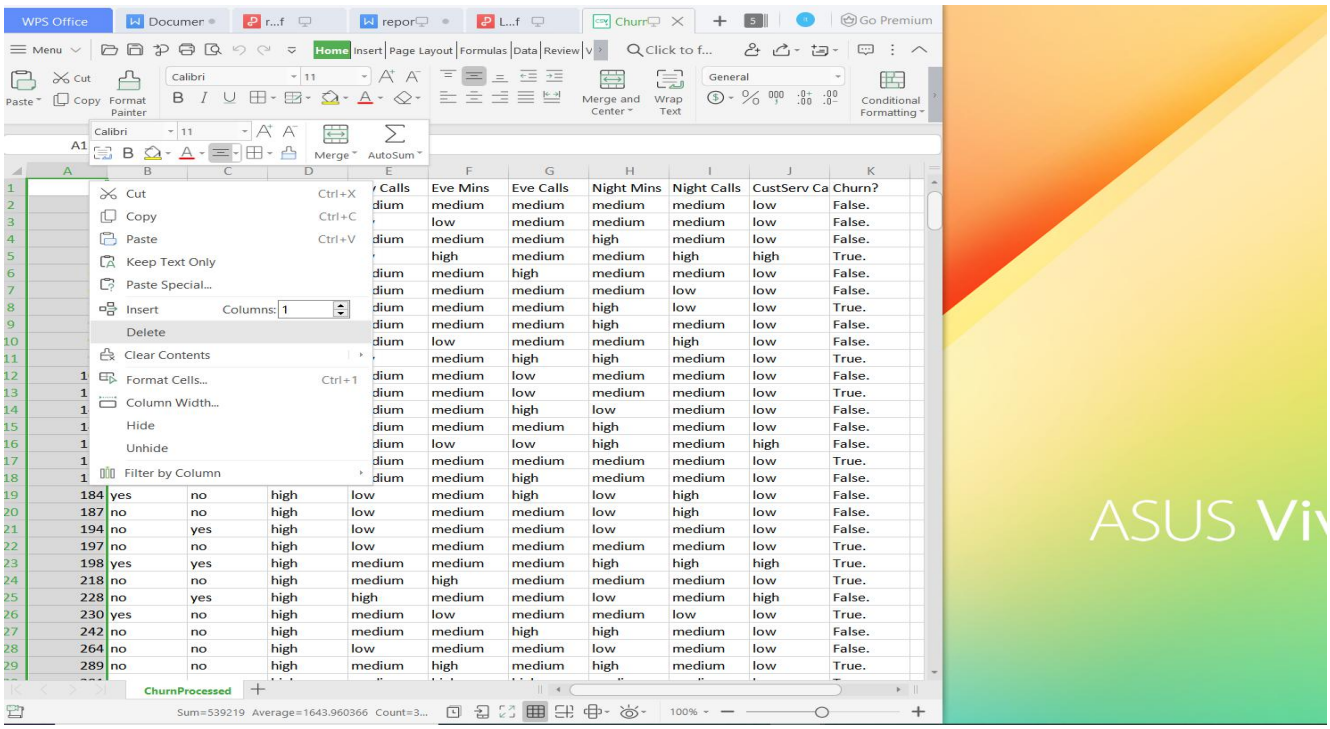
Các thuộc tính **Day Mins, Day Calls, Eve Mins, Eve Calls, Night Mins và Night Calls** đã được phân lớp.

- Lấy những dữ liệu có thuộc tính Day Mins =high.

	A	B	C	D	E	F	G	H	I	J	K
		Intl Plan	VMail Pla	Day Mins	Day Calls	Eve Mins	Eve Calls	Night Min	Night Call	CustServ	Churn?
2	0	no	yes	high	medium	medium	medium	medium	medium	low	False.
3	3	yes	no	high	low	low	medium	medium	medium	low	False.
4	9	yes	yes	high	medium	medium	medium	high	medium	low	False.
5	15	no	no	high	low	high	medium	medium	high	high	True.
6	61	no	yes	high	medium	medium	high	medium	medium	low	False.
7	66	yes	no	high	medium	medium	medium	medium	low	low	False.
8	76	no	no	high	medium	medium	medium	high	low	low	True.
9	93	no	no	high	medium	medium	medium	high	medium	low	False.
10	95	no	no	high	medium	low	medium	medium	high	low	False.
11	99	no	no	high	low	medium	high	high	medium	low	True.
12	106	no	yes	high	medium	medium	low	medium	medium	low	False.
13	117	no	no	high	medium	medium	low	medium	medium	low	True.
14	147	yes	no	high	medium	medium	high	low	medium	low	False.
15	149	no	yes	high	medium	medium	medium	high	medium	low	False.
16	154	no	no	high	medium	low	low	high	medium	high	False.
17	156	no	no	high	medium	medium	medium	medium	medium	low	True.
18	175	no	no	high	medium	medium	high	medium	medium	low	False.
19	184	yes	no	high	low	medium	high	low	high	low	False.

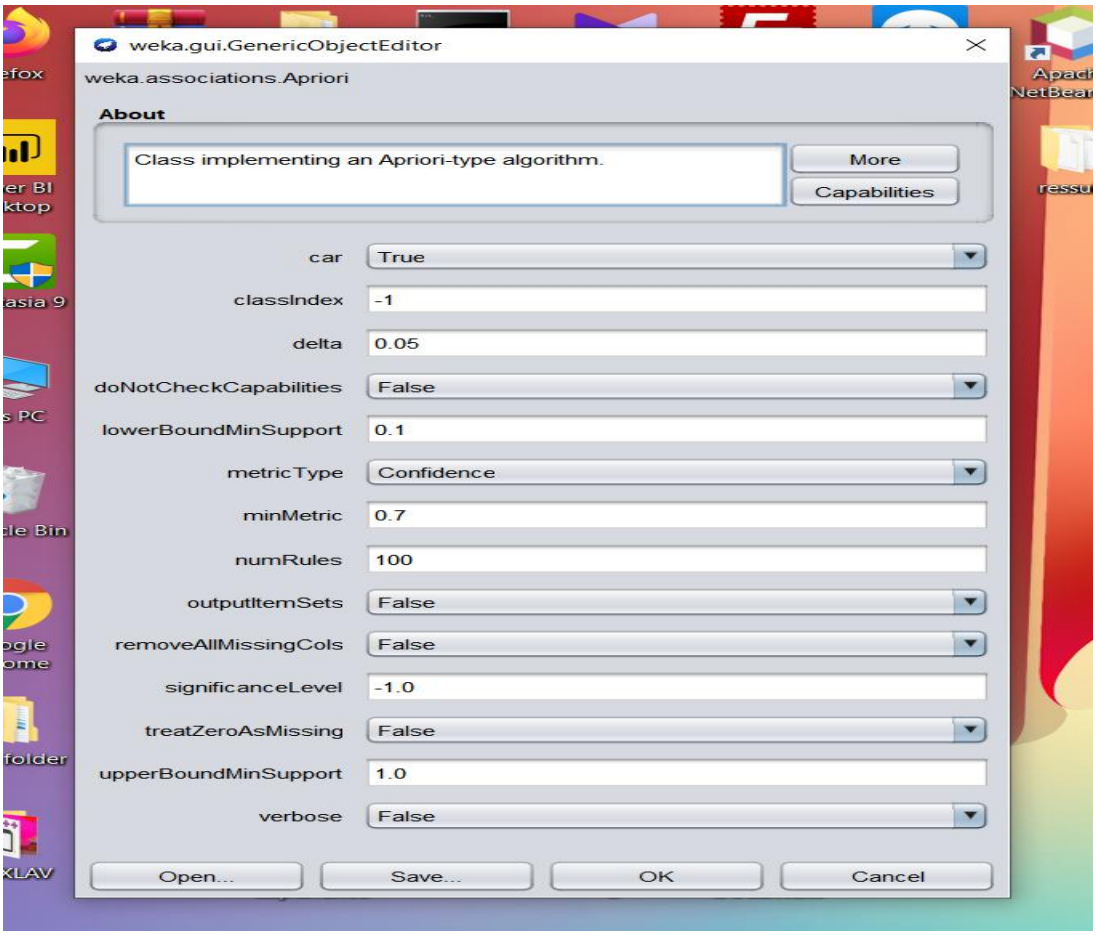
Thuộc tính Day Mins chỉ lấy các mẫu có **Day Mins= high**

⇨ Do sử dụng python để xuất ra file .csv nên ta cần xóa thuộc tính số thứ tự trong file ***ChurnProcessed.csv***



*Xóa cột đầu tiên trong file **ChurnProcessed.csv***

*** Điều chỉnh các tham số:**



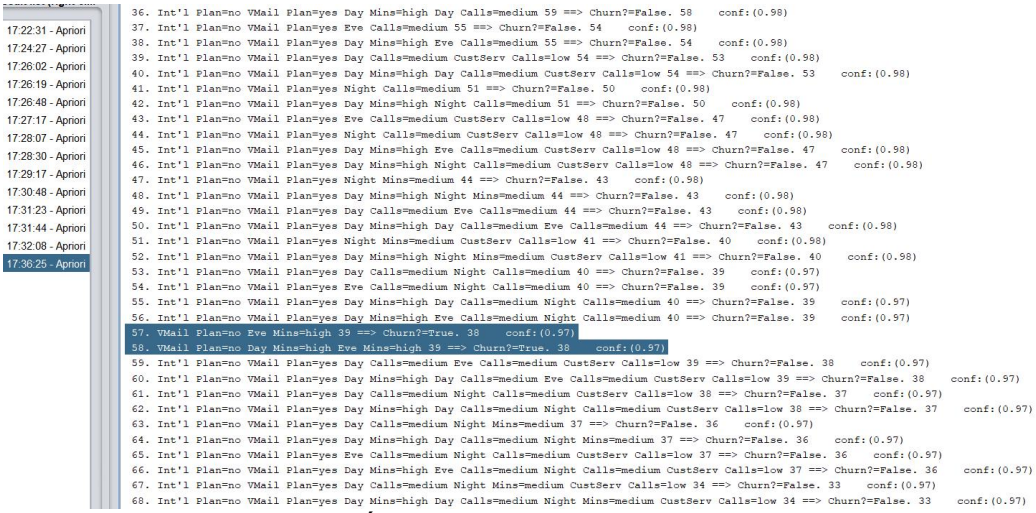
- Car: Chọn True để sinh ra các luật hướng về thuộc tính lớp (churn)
- numRules: số luật kỳ vọng sinh ra (trên thực tế nhều lúc numRules sinh ra thấp hơn kỳ vọng *(có thể là rất nhiều)*).

metricType: Loại độ đo được sử dụng: Confidence, lift, leverage, convinction. Ở đây mình dùng độ đo Confidence.
- lowBoundMinSupport: độ hỗ trợ nhỏ nhất.

***Kết quả phân tích:**

➢ Nhận xét quan trọng:

So với khi chưa xử lý với file process.py thì số luật thu được thấp hơn rất nhiều so với kỳ vọng.
➔ Sau khi trải qua bước xử lý thì số luật thu được đạt số lượng RẤT TỐT so với kỳ vọng:



Kết quả phân tích khi chạy với thuật toán Apriori

Dưới đây là các tập luật sinh ra Churn?=True cho Độ tin cậy(conf) cao nhất. (conf với Churn?=True giảm dần trong tập kết quả)

57. VMail Plan=no Eve Mins=high 39 ==> Churn?=True. 38 **conf:(0.97)**
58. VMail Plan=no Day Mins=high Eve Mins=high 39 ==> Churn?=True. 38 **conf:(0.97)**
135. VMail Plan=no Night Mins=high CustServ Calls=low 42 ==> Churn?=True. 36 **conf:(0.86)**
136. VMail Plan=no Day Mins=high Night Mins=high CustServ Calls=low 42 ==> Churn?=True. 36 **conf:(0.86)**

***Ý nghĩa**

Dựa vào kết quả trên đều cho thấy rằng. Nếu Số phút gọi ban đêm (Nigh Mins) và số phút gọi vào ban ngày (Day Mins) cao (high) thì khả năng rất cao là khách hàng này sẽ rời khỏi dịch vụ của công ty.
➔ ***Cần có những ưu đãi với những người có số phút gọi cao để giữ khách hàng ở lại.***

V. Tóm tắt kết quả:

1. Cách đánh giá kết quả:

Đánh giá dựa trên độ tin cậy của luật. Độ tin cậy được tính bằng cách: Giả sử ta có luật X-> Y thì độ tin cậy sẽ là:

$$confident = \frac{\sigma(X \cap Y)}{\sigma(X)}$$

Nếu độ tin cậy của luật phải lớn hơn độ tin cậy nhỏ nhất đã chỉ định từ trước thì nghĩa là kết quả có thể tin tưởng được:
confident* ≥ *minconfident

2. (Những) tập luật tốt nhất thu được:

57. VMail Plan=no Eve Mins=high 39 ==> Churn?=True. 38 conf:(0.97)
58. VMail Plan=no Day Mins=high Eve Mins=high 39 ==> Churn?=True. 38 conf:(0.97)
135. VMail Plan=no Night Mins=high CustServ Calls=low 42 ==> Churn?=True. 36 conf:(0.86)
136. VMail Plan=no Day Mins=high Night Mins=high CustServ Calls=low 42 ==> Churn?=True. 36 conf:(0.86)

***Ý nghĩa:**

- ✓ Luật **57**: Nếu không sử dụng dịch vụ Vmail và số phút gọi vào ban đêm cao thì người này rất có khả năng rời khỏi dịch vụ.
- ✓ Luật **58**: Nếu khách hàng này không dùng dịch vụ Vmail và có số phút gọi vào ban ngày và ban đêm cao thì khách hàng này khả năng cao sẽ từ bỏ sử dụng dịch vụ.
- ✓ Luật **135**: Nếu khách hàng này không dùng dịch vụ Vmail và có số phút gọi vào ban đêm và gọi đến dịch vụ CSKH cao thì khách hàng này khả năng cao sẽ từ bỏ sử dụng dịch vụ.
- ✓ Luật **136**: Nếu khách hàng này không dùng dịch vụ Vmail và có số phút gọi vào ban ngày&đêm và gọi đến dịch vụ CSKH cao thì khách hàng này khả năng cao sẽ từ bỏ sử dụng dịch vụ.

3. Ưu và nhược điểm trong bài tập:

a. Ưu điểm

- ✓ Có khả năng dùng python để mô phỏng các quan hệ của dữ liệu, biến đổi và chỉnh sửa dữ liệu.
- ✓ Sử dụng được thuật toán Apriori để phát sinh luật.
- ✓ Phát hiện ra thuộc tính lớp bị mất cân bằng và điều chỉnh bằng cách lấy một phần dữ liệu cân bằng để sinh luật.
- ✓ Phát hiệu và loại bỏ các thuộc tính "nhiễu".

b. Nhược điểm:

- 'Chưa thực sự' hiểu rõ bản chất toán học của các thuật toán sinh luật.
- Thời gian để hiểu rõ nội dung dữ liệu thực tế còn lâu.
- Đọc và 'hiểu' tài liệu tiếng anh chỉ mới dừng lại ở mức cơ bản.
- Chưa sinh ra đươc luật với các thuật toán quan trọng khác (nhưng vẫn cho ra kết qua cần hướng đến với Apriori).

VI. Nguồn ham khảo:

http://bis.net.vn/forums/p/384/675.aspx

https://github.com/IBM/telco-customer-churn-on-icp4d

https://medium.com/learning-intelligence/understanding-and-implementation-of-apriori-algorithm-with-python-part-2-ff037ceab254

Chuỗi bài giảng Khai phá dữ liệu: *http://ndhcuong.blogspot.com/p/khai-pha-du-lieu.html*

và nhiều nguồn khác

_ _ HẾT _ _