

**Đại học Quốc gia TP.HCM**  
**Trường Đại học Khoa học Tự nhiên**

---

**KHAI THÁC TẬP PHỔ BIẾN  
& LUẬT KẾT HỢP**

Khai thác dữ liệu & Ứng dụng

Nguyễn Bảo Long - MSSV: 18120201

Huỳnh Long Nam - MSSV: 18120212

---

TP Hồ Chí Minh, ngày 14/11/2020

# Contents

<b>1</b>	<b>Thông tin chung</b>	<b>2</b>
<b>2</b>	<b>Tìm hiểu về tập dữ liệu</b>	<b>3</b>
<b>3</b>	<b>Tiền xử lý</b>	<b>3</b>
3.1	Loại bỏ các thuộc tính có tương quan với nhau . . . . .	3
3.2	Loại bỏ các thuộc tính dị thường & thuộc tính index . . . . .	9
<b>4</b>	<b>Khám phá dữ liệu</b>	<b>10</b>
4.1	Khám phá dữ liệu dạng nominal . . . . .	10
4.2	Khám phá dữ liệu dạng numeric . . . . .	11
4.3	Tương quan đa thuộc tính với numeric . . . . .	13
<b>5</b>	<b>Luật kết hợp</b>	<b>15</b>
<b>6</b>	<b>Kết luận</b>	<b>18</b>

# 1 Thông tin chung

## 1. Thông tin thành viên nhóm

STT	Họ tên	MSSV	Email
1	Nguyễn Bảo Long	18120201	18120201@student.hcmus.edu.vn
2	Huỳnh Nam Long	18120212	18120212@student.hcmus.edu.vn

Table 1: Bảng thông tin thành viên nhóm

## 2. Tỷ lệ tham gia công việc

STT	MSSV	Công việc	Hoàn thành
1	18120201	Rút trích luật từ tập dữ liệu	Đã hoàn thành
2		Trình bày báo cáo	Đã hoàn thành
3	18120212	Tìm hiểu ý nghĩa tập dữ liệu	Đã hoàn thành
3		Tiền xử lý dữ liệu	Đã hoàn thành

Table 2: Bảng phân chia công việc

## 2 Tìm hiểu về tập dữ liệu

- Số lượng mẫu: 3333
- Số lượng thuộc tính: 21
- Thuộc tính lớp: Churn
- Ý nghĩa của thuộc tính
  1. State: tên các bang, quận ở Columbia (nominal)
  2. Account length: thời gian kích hoạt của tài khoản (numeric)
  3. Area code: mã vùng (nominal)
  4. Phone number: số điện thoại, có thể thay thế cho ID khách hàng (numeric)
  5. International Plan: dịch vụ cuộc gọi quốc tế (nominal, gồm 2 giá trị Yes và No)
  6. VoiceMail Plan: dịch vụ thư thoại (nominal, gồm 2 giá trị Yes và No)
  7. Number of voice mail messages: số tin nhắn thoại (nominal)
  8. Total day minutes: tổng thời gian khách hàng sử dụng vào ban ngày
  9. Total day calls: tổng số cuộc gọi vào ban ngày
  10. Total day charge: tổng cước vào ban ngày, dựa vào 2 biến trên
  11. Total evening minutes: tổng thời gian sử dụng vào buổi tối
  12. Total evening calls: tổng số cuộc gọi vào buổi tối
  13. Total evening charge: tổng cước vào buổi tối, dựa vào 2 biến trên
  14. Total night minutes: tổng thời gian sử dụng vào ban đêm
  15. Total night calls: tổng số cuộc gọi vào ban đêm
  16. Total night charge: tổng cước vào ban đêm, dựa vào 2 biến trên
  17. Total international minutes: tổng thời gian gọi quốc tế
  18. Total international calls: tổng số cuộc gọi quốc tế
  19. Total international charge: tổng cước quốc tế, dựa vào 2 biến trên
  20. Number of calls to customer service: số cuộc gọi chăm sóc khách hàng
  21. Churn: Có hủy dịch vụ hay không
- Để hiểu tập dữ liệu rõ hơn, cần phải xác định các thuộc tính có ảnh hưởng lớn đến sự phân lớp của thuộc tính lớp ‘Churn?’

## 3 Tiền xử lý

### 3.1 Loại bỏ các thuộc tính có tương quan với nhau

- Lý do loại bỏ: Nếu không loại bỏ các thuộc tính này, dữ liệu sẽ bị nhấn mạnh quá mức. Trong trường hợp xấu nhất, model có thể không ổn định và cung cấp kết quả không đáng tin cậy
- Phương pháp:

- Sử dụng tab *Visualize* để dự đoán các thuộc tính có tương quan tuyến tính với nhau theo 1 hàm nào đó

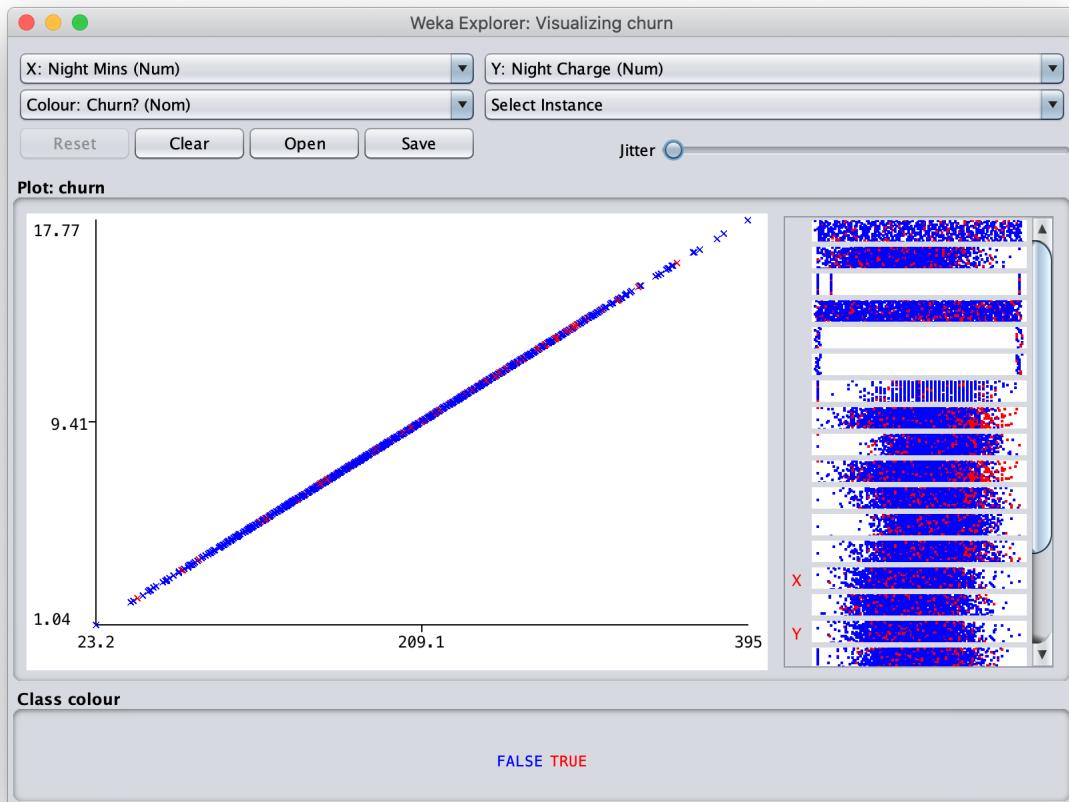


Figure 1: Biểu diễn quan hệ giữa Night Mins (X) và Night Charge (Y)

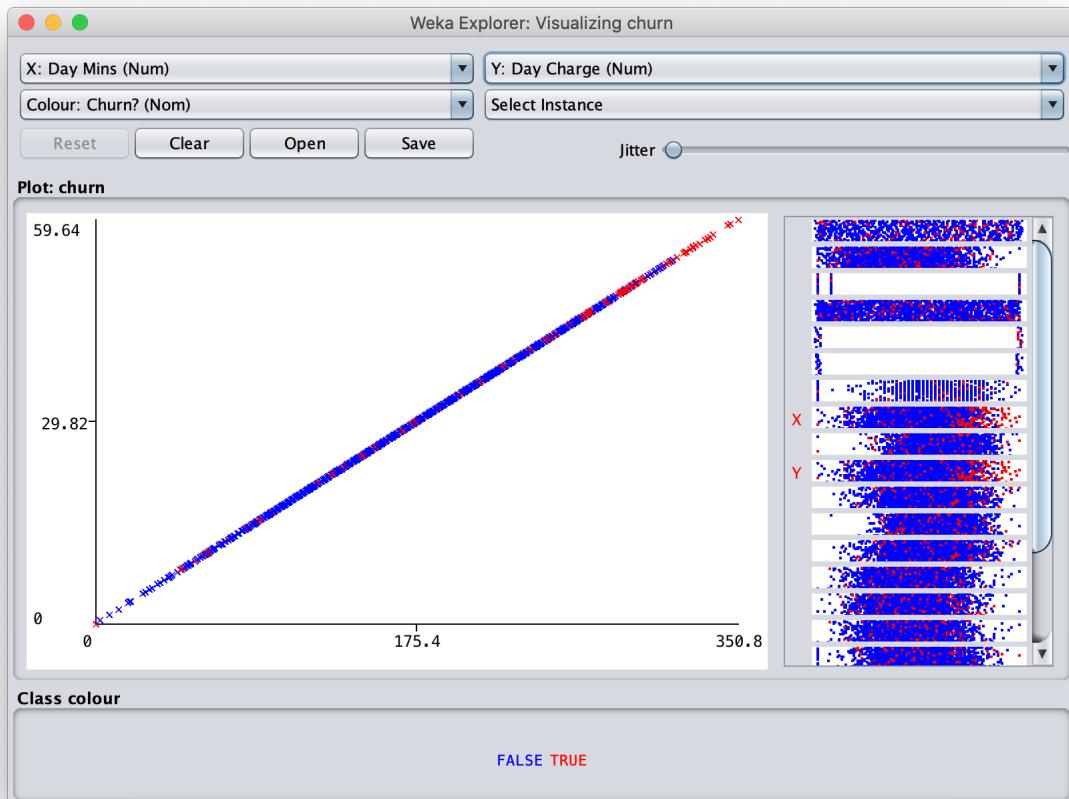


Figure 2: Biểu diễn quan hệ giữa Day Mins (X) và Day Charge (Y)

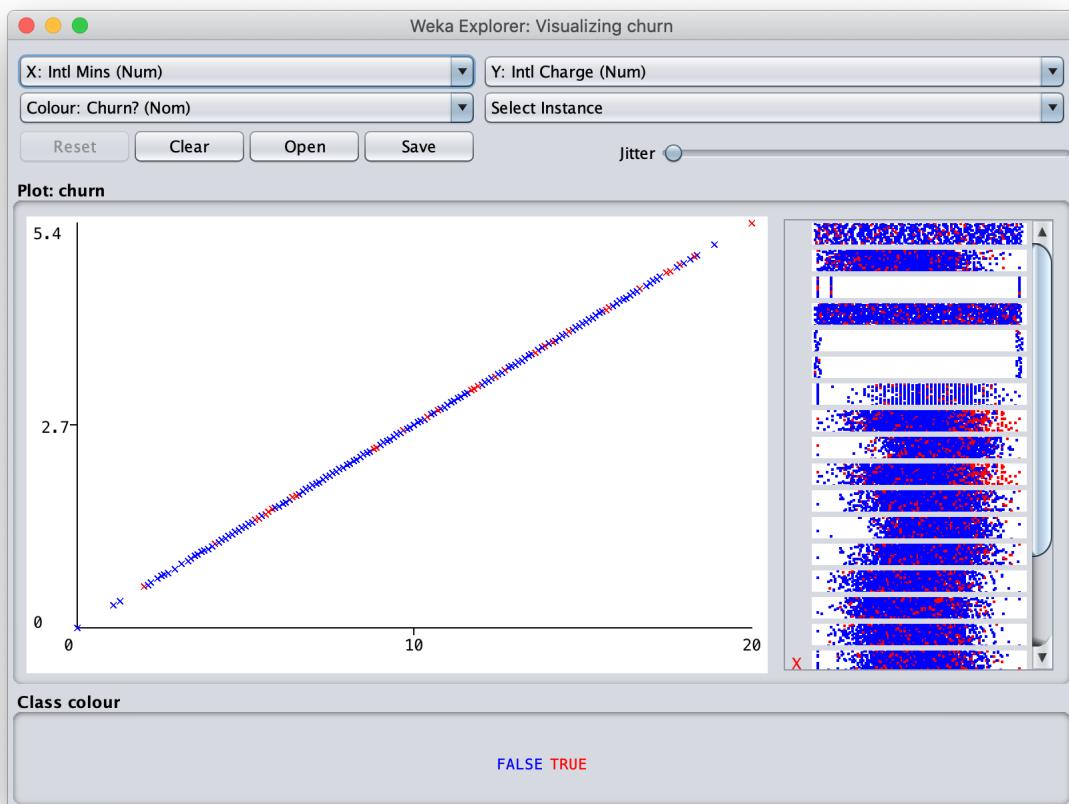


Figure 3: Biểu diễn quan hệ giữa Intl Mins (X) và Intl Charge (Y)

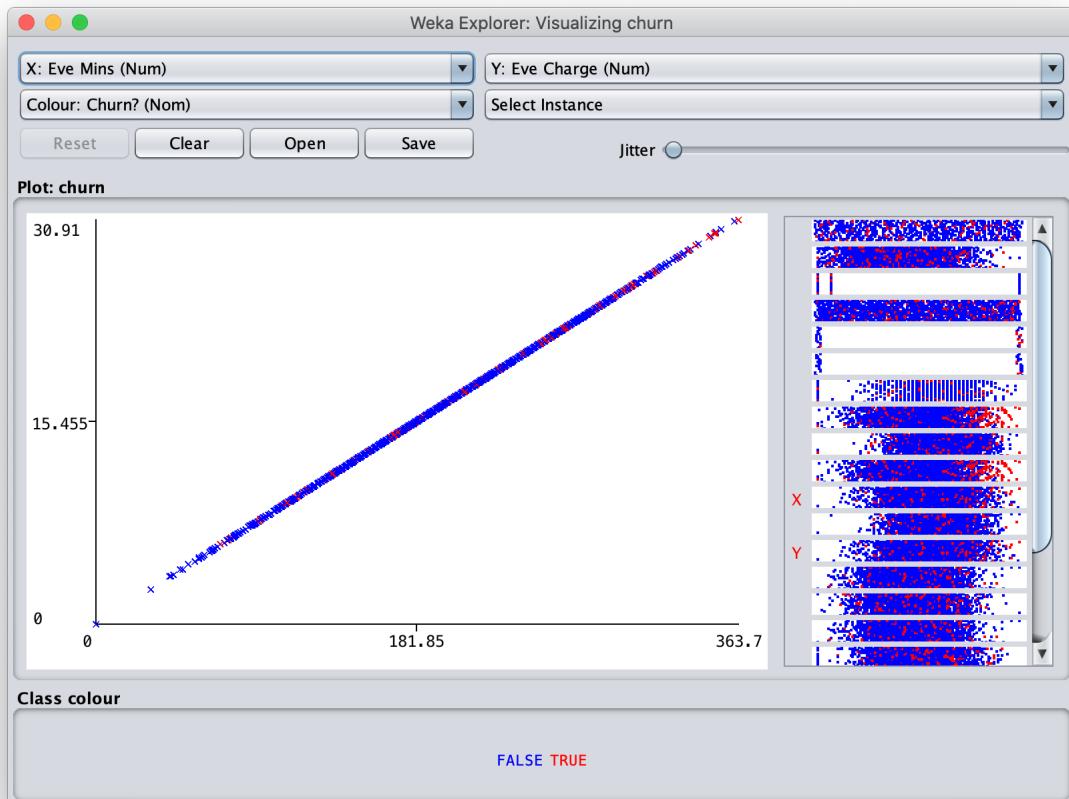


Figure 4: Biểu diễn quan hệ giữa Eve Mins (X) và Eve Charge (Y)

- Từ 3 hình trên, ta có thể dự đoán:
  - \* Night Mins (X) và Night Charge (Y) có tương quan tuyến tính với nhau
  - \* Day Mins (X) và Day Charge (Y) có tương quan tuyến tính với nhau
  - \* Intl Mins (X) và Intl Charge (Y) có tương quan tuyến tính với nhau
  - \* Eve Mins (X) và Eve Charge (Y) có tương quan tuyến tính với nhau
- Sử dụng tab *Classify* để tìm ra hàm tuyến tính biểu diễn mối tương quan này
  - \* B1: Tại tab *Preprocess*, remove các thuộc tính không cần xác định mỗi tương quan. Ví dụ khi muốn xác định mối tương quan giữa Night Mins và Night Charge, cần remove hết các thuộc tính khác chỉ để lại 2 thuộc tính này.
  - \* B2: Tại tab *Classify*, chọn Classifier là *Linear Regression*; sau đó chọn thuộc tính phụ rồi bấm nút *Start* và đọc kết quả thu được
- Sau khi đã xác định quan hệ giữa các thuộc tính được dự đoán bên trên, ta có các kết quả sau

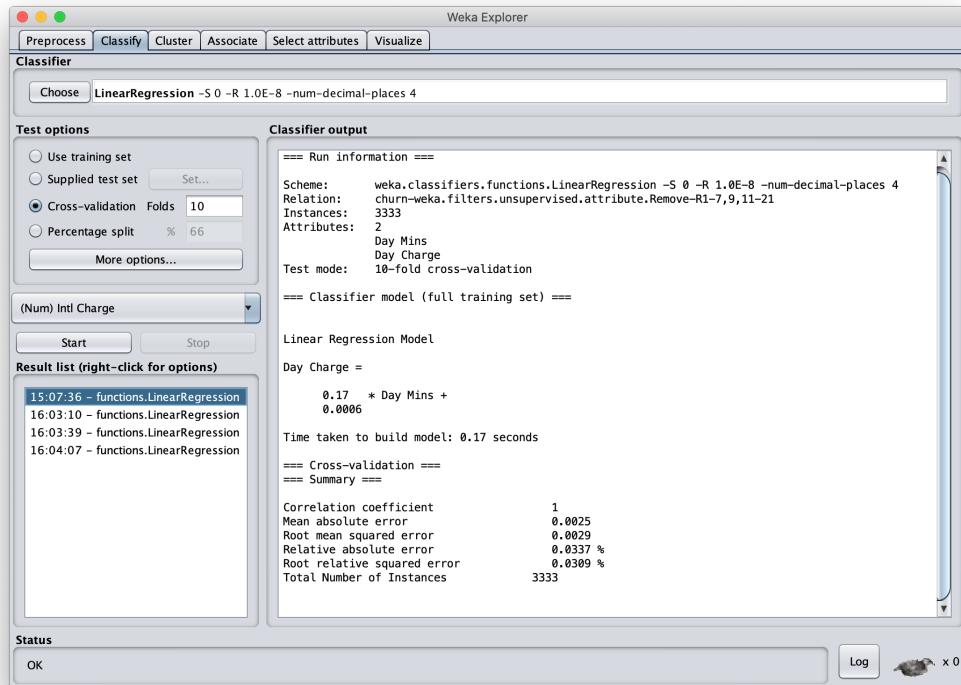


Figure 5: Mối tương quan tuyến tính giữa Day Mins và Day Charge

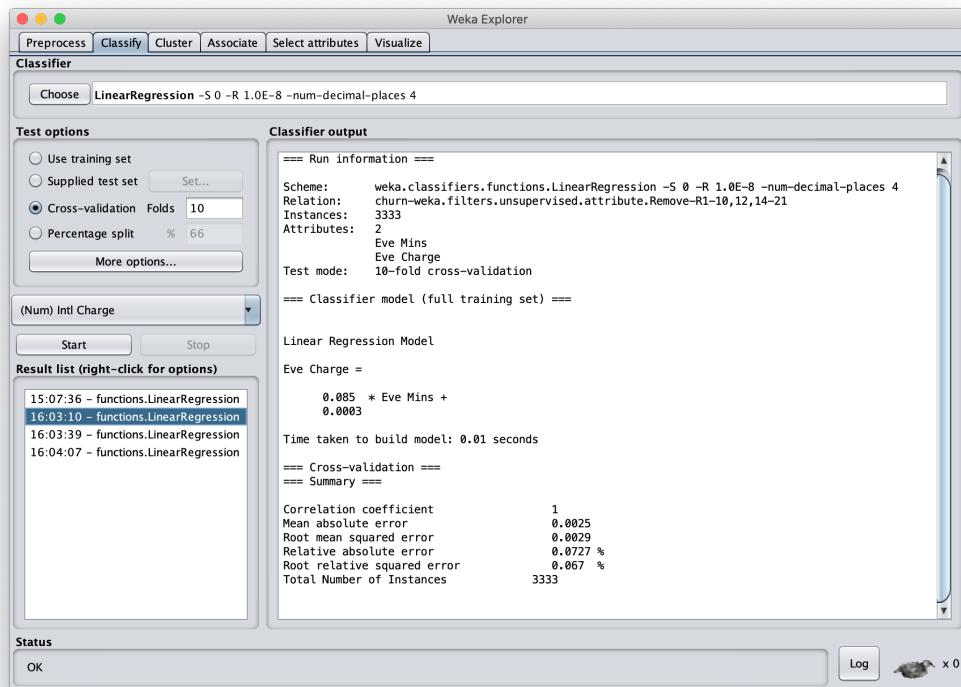


Figure 6: Mối tương quan tuyến tính giữa Eve Mins và Eve Charge

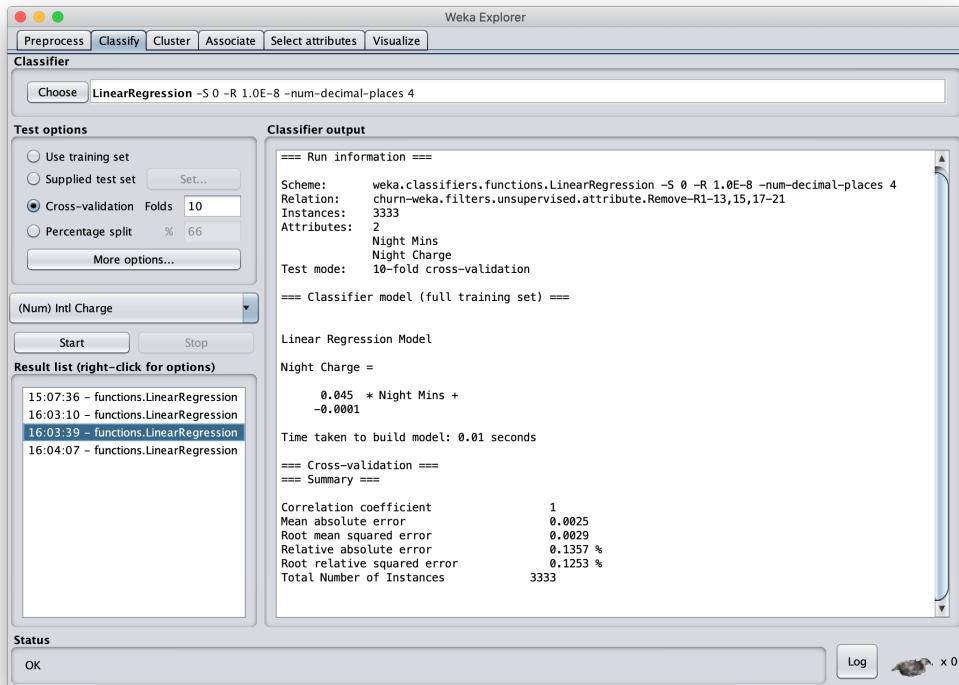


Figure 7: Mối tương quan tuyến tính giữa Night Mins và Night Charge

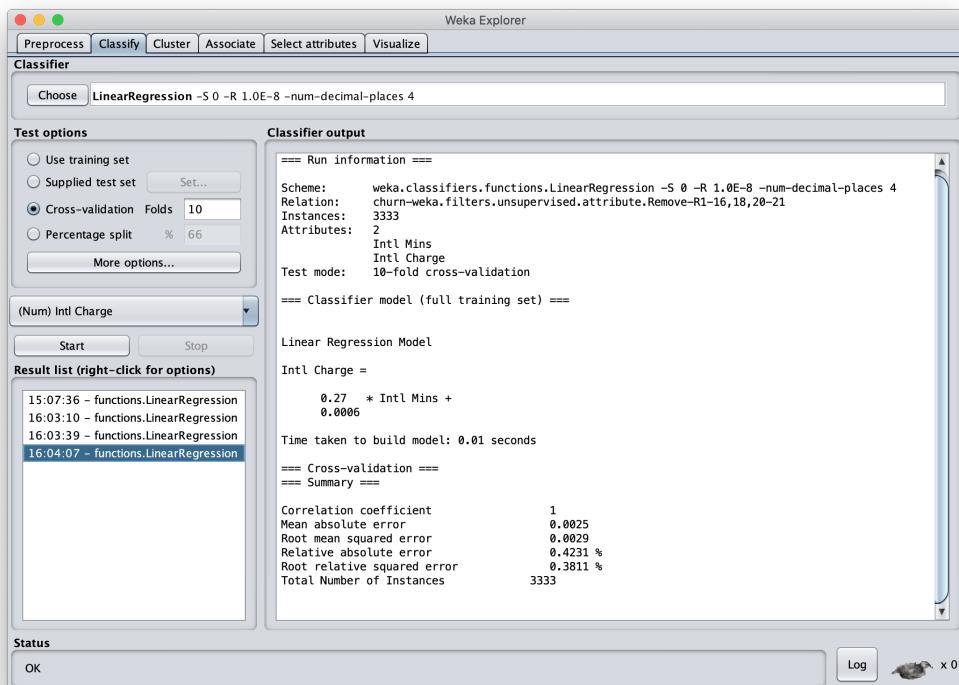


Figure 8: Mối tương quan tuyến tính giữa Intl Mins và Intl Charge

- Từ kết quả trên, ta có thể bỏ các thuộc tính phụ thuộc *Eve Charge*, *Day Charge*, *Night Charge*, *Intl Charge*

### 3.2 Loại bỏ các thuộc tính dị thường & thuộc tính index

- Đối với thuộc tính dị thường: Thuộc tính *State* chứa tên viết tắt của 51 bang nước Mỹ. Nhưng thuộc tính *Area Code* thì chỉ có 3 giá trị: 408, 415, 510 đều là *Area Code* thuộc California. Theo đó, tất cả các khách hàng có trong tập dữ liệu đều sống ở California. Chính điều này gây ra sự ‘dị thường’
- Đối với thuộc tính index: Trong dataset có thuộc tính *Phone* có thể thay thế cho ID của khách hàng
- Cách xử lý cho các thuộc tính này
  - Xoá thuộc tính
  - Tham khảo người tạo ra tập dữ liệu để hiểu thêm về thuộc tính này

## 4 Khám phá dữ liệu

### 4.1 Khám phá dữ liệu dạng nominal

- Đối với thuộc tính *International Plan*

- Nhập vào tên thuộc tính để xác định phần trăm của thuộc tính class tương ứng với mỗi lựa chọn
- Trong các mẫu *International Plan = No* (3010 mẫu), có 11.5% huỷ dịch vụ
- Trong các mẫu *International Plan = Yes* (323 mẫu), có 42.4% huỷ dịch vụ
- Do đó có thể thấy: khách hàng có *International Plan = Yes* có tỷ lệ huỷ dịch vụ cao hơn khách hàng có *International Plan = No*

```

nguyenbaolong@MBP-cua-Nguyen:~/Desktop/Không phải nhập/DataMining...
nguyenbaolong@MBP-cua-Nguyen > ~/Desktop/Không phải nhập/DataMining-Lab2/Source >
main + python code.py
Enter attribute: Intl Plan
[11.495016611295682, 3010, 'no']
[42.414860681114554, 323, 'yes']
nguyenbaolong@MBP-cua-Nguyen > ~/Desktop/Không phải nhập/DataMining-Lab2/Source >
main + 

```

Figure 9: Tỷ lệ huỷ dịch vụ theo thuộc tính International Plan

- Đối với thuộc tính *Voice Mail Plan*

- Nhập vào tên thuộc tính để xác định phần trăm của thuộc tính class tương ứng với mỗi lựa chọn
- Trong các mẫu *Voice Mail Plan = No* (2411 mẫu), có 16.7% huỷ dịch vụ
- Trong các mẫu *Voice Mail Plan = Yes* (922 mẫu), có 8.6% huỷ dịch vụ
- Do đó có thể thấy: khách hàng có *Voice Mail Plan = No* có tỷ lệ huỷ dịch vụ cao hơn khách hàng có *Voice Mail Plan = Yes*

```

nguyenbaolong@MBP-cua-Nguyen:~/Desktop/Không phải nhập/DataMining...
nguyenbaolong@MBP-cua-Nguyen > ~/Desktop/Không phải nhập/DataMining-Lab2/Source
main + python code.py
Enter attribute: VMail Plan
[16.71505599336375, 2411, 'no']
[8.676789587852495, 922, 'yes']
nguyenbaolong@MBP-cua-Nguyen > ~/Desktop/Không phải nhập/DataMining-Lab2/Source
main +

```

Figure 10: Tỷ lệ huỷ dịch vụ theo thuộc tính Voice Mail Plan

## 4.2 Khám phá dữ liệu dạng numeric

- Dối với thuộc tính *Account Length*

- Nhập vào tên thuộc tính và số khoảng chia để rời rạc hoá dữ liệu của thuộc tính
- Đọc dữ liệu cho kết quả đầu tiên [7.4074074074074066, 27, 1.0, 10.0]: Trong 27 mẫu có giá trị trong khoảng [1.0, 10.0] có 7.4% số mẫu có ‘Churn?’ = True
- Quan sát tỷ lệ phần trăm, ta nhận thấy tỷ lệ khá đồng đều. Do đó không có điều gì đặc biệt có thể tác động đến sự phân lớp ở trong thuộc tính này
- Việc tương tự xảy ra khi chạy chương trình với các thuộc tính *Day Calls*, *Evening Calls*, *Night Calls*, *International Calls*, *Night Mins*, *International Mins*, *Voice Mail Message*

```

main + python code.py
Enter attribute: Account Length
Num of bin: 25
3333
[7.4074074074066, 27, 1.0, 10.0]
[14.634146341463413, 41, 11.0, 20.0]
[16.666666666666664, 60, 21.0, 30.0]
[9.63855421686747, 83, 31.0, 39.0]
[13.91304347826087, 115, 40.0, 49.0]
[12.5, 168, 50.0, 59.0]
[14.622641509433961, 212, 60.0, 68.0]
[15.769230769230768, 260, 69.0, 78.0]
[11.627906976744185, 301, 79.0, 88.0]
[15.753424657534246, 292, 89.0, 97.0]
[14.035087719298245, 342, 98.0, 107.0]
[17.16171617161716, 303, 108.0, 117.0]
[16.417910447761194, 268, 118.0, 126.0]
[15.040650406504067, 246, 127.0, 136.0]
[11.475409836065573, 183, 137.0, 146.0]
[14.705882352941178, 136, 147.0, 155.0]
[13.274336283185843, 113, 156.0, 165.0]
[18.181818181818183, 66, 166.0, 175.0]
[18.367346938775512, 49, 176.0, 184.0]
[6.0606060606060606, 33, 185.0, 194.0]
[13.333333333333334, 15, 195.0, 204.0]
[30.0, 10, 205.0, 212.0]
[0.0, 4, 215.0, 221.0]
[33.3333333333333, 6, 224.0, 243.0]

```

Figure 11: Tỷ lệ huỷ dịch vụ theo thuộc tính Account Plan

- Đối với thuộc tính *Day Mins*
  - Nhập vào tên thuộc tính và số khoảng chia để rời rạc hoá dữ liệu của thuộc tính
  - Đọc dữ liệu cho kết quả đầu tiên [16.666666666666664, 6, 0.0, 12.5]: Trong 6 mẫu có giá trị trong khoảng [0.0, 12.5] có 16.6% số mẫu có ‘Churn?’ = True
  - Quan sát tỷ lệ phần trăm, ta nhận thấy tỷ lệ huỷ dịch vụ tăng khi Day Mins tăng. Như vậy rõ ràng thuộc tính này có thể dùng để dự đoán kết quả phân lớp ‘Churn?’
  - Việc tương tự xảy ra khi chạy chương trình với các thuộc tính *Eve Mins*, *Customer Service Calls*

```
nguyenbaolong@MBP-cua-Nguyen:~/Desktop/Không phải nháp/DataMining...
main + python code.py
Enter attribute: Day Mins
Num of bin: 25
3333
[16.66666666666666, 6, 0.0, 12.5]
[0.0, 5, 17.6, 27.0]
[0.0, 10, 29.9, 41.9]
[18.181818181818183, 22, 44.9, 55.6]
[13.157894736842104, 38, 57.1, 69.4]
[13.461538461538462, 52, 70.7, 83.8]
[10.416666666666668, 96, 84.2, 98.2]
[11.811023622047244, 127, 98.4, 112.2]
[12.994350282485875, 177, 112.6, 126.1]
[14.473684210526317, 228, 126.3, 140.2]
[8.945686900958465, 313, 140.4, 154.3]
[13.190184049079754, 326, 154.4, 168.3]
[8.206686930091186, 329, 168.4, 182.3]
[5.214723926380368, 326, 182.5, 196.4]
[5.379746835443038, 316, 196.5, 210.4]
[8.620689655172415, 290, 210.5, 224.5]
[20.28985507246377, 207, 224.6, 238.4]
[23.89937106918239, 159, 238.8, 252.4]
[27.67857142857143, 112, 252.6, 266.6]
[53.33333333333336, 90, 266.7, 280.4]
[57.446808510638306, 47, 281.0, 294.2]
[60.71428571428571, 28, 294.7, 308.6]
[76.47058823529412, 17, 309.9, 322.5]
[100.0, 8, 324.7, 335.5]
[100.0, 4, 337.4, 350.8]
nguyenbaolong@MBP-cua-Nguyen > ~/Desktop/Không phải nháp/DataMining-Lab2/Source▶ h
main +
```

Figure 12: Tỷ lệ huỷ dịch vụ theo thuộc tính Day Mins

### 4.3 Tương quan đa thuộc tính với numeric

- Service Call cao và Day Mins thấp thì tỷ lệ huỷ dịch vụ cao
- Day Mins cao thì tỷ lệ huỷ dịch vụ cao

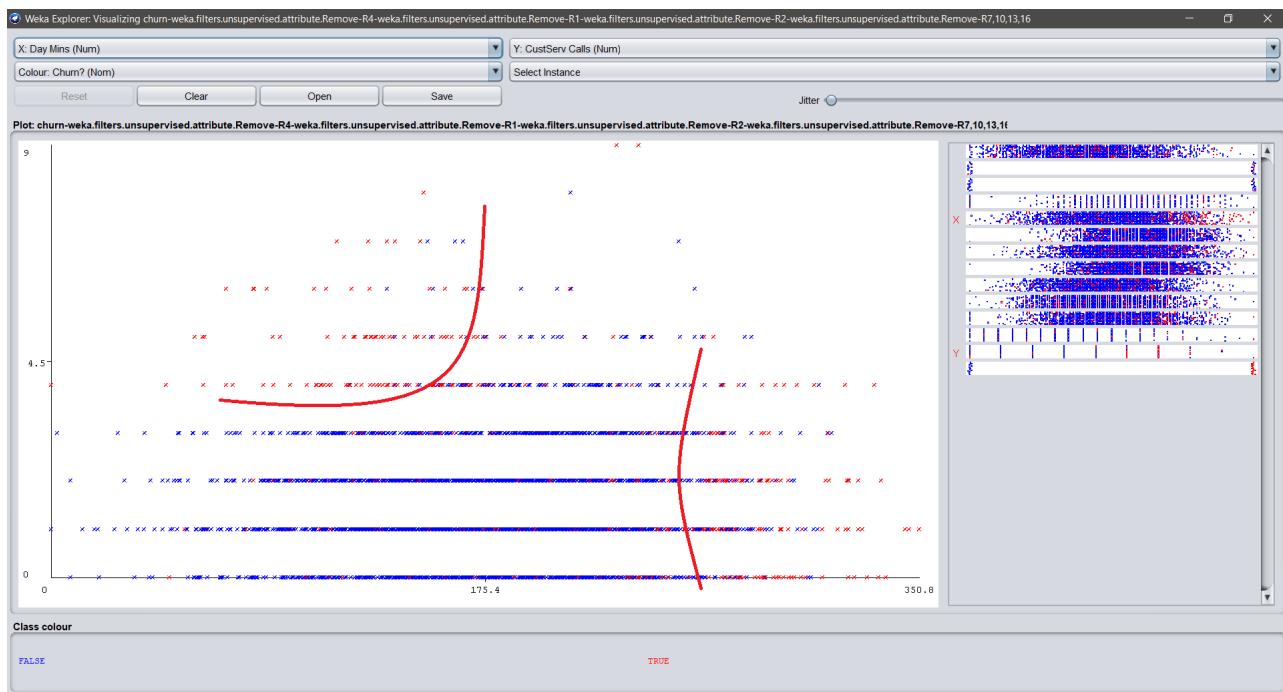


Figure 13: Tương quan đa thuộc tính giữa Service Call và Day Mins

## 5 Luật kết hợp

- Phương pháp đánh giá luật: Việc đánh giá luật được thể hiện qua 2 thông số *support*, *confidence*
  - Độ *support* (độ hỗ trợ) đo tần số xuất hiện của phần tử hay tập phần tử. *Min support* (ngưỡng hỗ trợ tối thiểu) là giá trị *support* nhỏ nhất được chỉ định bởi người dùng
 
$$support(A \Rightarrow B) = P(A \cup B)$$
  - Độ *confidence* (độ tin cậy) đo tần số xuất hiện của phần tử hay tập phần tử trong điều kiện xuất hiện của phần tử hay tập phần tử khác. *Min confidence* (ngưỡng tin cậy tối thiểu) là giá trị *confidence* nhỏ nhất được chỉ định bởi người dùng
 
$$confidence(A \Rightarrow B) = P(B|A)$$

- Chọn tab *Preprocess* để rời rạc hoá dữ liệu bằng cách chọn filter *Discretize* (trong mục *./filters/unsupervised/Discretize*) và tùy chỉnh các thông số của bộ lọc như hình



Figure 14: Tùy chỉnh các thông số cho quá trình rời rạc hoá dữ liệu

- Chọn tab *Associate*, thuật toán *Apriori* để tiến hành rút trích luật
- Tùy chỉnh các thông số *car*, *lowerBoundMinSupport*, *metricType*, *minMetric*, *numRules*, *upperBoundMinSupport* như hình để có thể lấy được các luật có **Churn? = TRUE**



Figure 15: Tuỳ chỉnh các thông số cho thuật toán Apriori

- Kết quả thu được 2842 luật (file luật được đính kèm trong folder Data). Trong đó, có một vài luật rất đáng chú ý khi so sánh với các dự đoán khi phân tích các thuộc tính được trình bày ở các mục trên
  - **Rule 6:**  $\text{Intl Plan=yes Intl Calls='(1.5-2.5]'} \ 47 ==> \text{Churn?=TRUE} \ 47 \text{ conf:}(1)$
  - **Rule 51:**  $\text{Intl Plan=yes VMail Plan=no Intl Calls='(1.5-2.5]'} \ 35 ==> \text{Churn?=TRUE} \ 35 \text{ conf:}(1)$
  - **Rule 52:**  $\text{Intl Plan=yes VMail Message='(-inf-2]'} \ \text{Intl Calls='(1.5-2.5]'} \ 35 ==> \text{Churn?=TRUE} \ 35 \text{ conf:}(1)$
  - **Rule 2144:**  $\text{VMail Plan=no Day Mins='(274.8-inf)'} \ 98 ==> \text{Churn?=TRUE} \ 82 \text{ conf:}(0.84)$
  - **Rule 2146:**  $\text{VMail Message='(-inf-2]'} \ \text{Day Mins='(274.8-inf)'} \ 98 ==> \text{Churn?=TRUE} \ 82 \text{ conf:}(0.84)$
  - **Rule: 2752:**  $\text{VMail Plan=no CustServ Calls='(3.5-4.5]'} \ 124 ==> \text{Churn?=TRUE} \ 60 \text{ conf:}(0.48)$
  - **Rule: 2753:**  $\text{VMail Message='(-inf-2]'} \ \text{CustServ Calls='(3.5-4.5]'} \ 124 ==> \text{Churn?=TRUE} \ 60 \text{ conf:}(0.48)$
  - **Rule: 2754:**  $\text{VMail Plan=no VMail Message='(-inf-2]'} \ \text{CustServ Calls='(3.5-4.5]'} \ 124 ==> \text{Churn?=TRUE} \ 60 \text{ conf:}(0.48)$

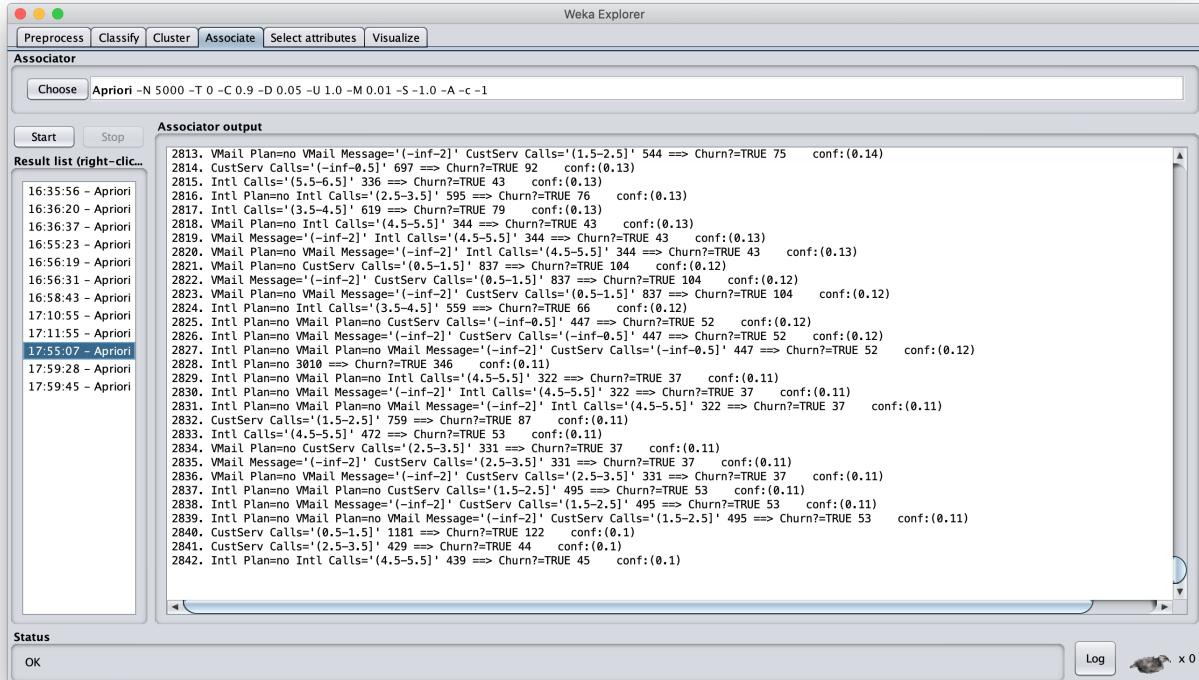


Figure 16: Rút trích luật từ tập dữ liệu

- Nhận xét luật

- Các luật 6, 51, 52 có thuộc tính *Intl Plan = yes* hoặc *VMail Plan = no* và có ảnh hưởng lớn đến kết quả phân lớp như đã dự đoán
- Các luật 2144, 2146 có thuộc tính *Day Mins = '(274.8-inf)'* (Day Mins lớn) và có ảnh hưởng lớn đến kết quả phân lớp như đã dự đoán
- Các luật 2752, 2753, 2754 có thuộc tính *CustServ Calls='(3.5-4.5)'* (CustServ Calls cao) và có ảnh hưởng lớn đến kết quả phân lớp

## 6 Kết luận

Từ những ý phân tích trên, ta có thể rút ra 1 vài kết luận như sau

- Khách hàng có đăng ký International Plan thì sẽ có tỷ lệ huỷ dịch vụ sẽ cao hơn khách hàng không đăng ký
- Customer Service càng cao thì khả năng từ chối dịch vụ càng cao
- Khi Eve mins tăng thì tỷ lệ từ chối dịch vụ tăng nhẹ (không ảnh hưởng nhiều đến phân lớp)
- Service Call tăng thì tỷ lệ từ chối dịch vụ tăng mạnh
- Service Call cao và Day Mins thấp thì tỷ lệ huỷ dịch vụ cao
- Day Mins cao thì tỷ lệ huỷ dịch vụ cao