Data Scientist Testing - FPT

Lanh Nguyen

Ngày 3 tháng 11 năm 2017

Bài toán

Tập dataset mô tả lịch sử sử dụng dịch vụ truyền hình HD của các user đã hủy hợp đồng hệ thống. Sử dụng ngôn ngữ lập trình (java, python, R, dotnet...) thực hiện các công việc sau:

- Parse log lấy ra các thuộc tính:
  - `MAC`: userID
  - `SessionMainMenu`: thời điểm user bắt đầu sử dụng dịch vụ
  - `AppName`: loại app user sử dụng
  - `LogID`: mã log
  - `Event`: thao tác user
  - `ItemID`: ID chương trình user xem
  - `RealTimePlaying`: thời lượng xem của user.

  → Output: tập log mới với format là dạng row, column. Mỗi row tương ứng dòng log, mỗi column tương ứng trường thuộc tính. Mỗi column cách nhau dấu tab
- Kết hợp file user_info.txt và tập data đã parse, phân tích hành vi đặc điểm sử dụng dịch vụ của những user này
- Dựa vào kết quả phân tích, đề ra giải pháp dự đoán user có khả năng hủy sử dụng dịch vụ

1 Parse log file

Source Code: Github, Jupyter NooteBook.

Thống kê Tổng số mẫu: 914060 entries

Bảng 1: Thống kê tập dữ liệu đã trích xuất.

No	Feature	Type	Ratio missing
1	`MAC`	string	0/914060
2	`SessionMainMenu`	string	28/914060
3	`AppName`	category	0/914060
4	`LogID`	numeric	0/914060
5	`Event`	category	0/914060
6	`ItemID`	category	153122/914060
7	`RealTimePlaying`	numeric	620029/914060

Data preprocessing with missing values

- Đối với dữ liệu dạng số: `RealTimeplaying` → thay thế bởi 0.
- Đối với dữ liệu dạng chuỗi: `SessionMainMenu`, và `ItemID` → thay thế bởi chuỗi 'NA'.

Bảng dữ liệu data log Hình 1 trích xuất 10 mẫu từ tập dữ liệu sau khi đã tiền xử lý.

	MAC	SessionMainMenu	AppName	LogID	Event	ItemID	RealTimePlaying
0	B046FCAC0DC1	B046FCAC0DC1:2016:02:12:12:35:13:437	VOD	52	StopVOD	100052388	570.3
1	B046FCAC0DC1	B046FCAC0DC1:2016:02:11:01:01:56:838	IPTV	40	EnterIPTV	NA	0.0
2	B046FCAC0DC1	B046FCAC0DC1:2016:02:11:01:02:29:258	VOD	55	NextVOD	100052388	0.0
3	B046FCAC0DC1	B046FCAC0DC1:2016:02:12:04:44:59:143	IPTV	18	ChangeModule	NA	0.0
4	B046FCAC0DC1	B046FCAC0DC1:2016:02:12:12:35:13:437	VOD	54	PlayVOD	100052388	0.0
5	B046FCAC0DC1	B046FCAC0DC1:2016:02:12:04:44:59:143	IPTV	40	EnterIPTV	NA	0.0
6	B046FCAC0DC1	B046FCAC0DC1:2016:02:12:12:35:13:437	VOD	55	NextVOD	100052388	0.0
7	B046FCAC0DC1	B046FCAC0DC1:2016:02:12:12:35:13:437	VOD	52	StopVOD	100052388	3384.6
8	B046FCAC0DC1	B046FCAC0DC1:2016:02:13:17:25:40:373	IPTV	40	EnterIPTV	NA	0.0
9	B046FCAC0DC1	B046FCAC0DC1:2016:02:14:01:41:40:431	VOD	52	StopVOD	100052388	621.9

Hình 1: Bảng dữ liệu sau khi đã tiền xử lý

Dữ liệu user-info. Kết hợp nhiều racord có cùng user id lại với nhau bằng cách tính tổng thời lượng sử dụng ứng dụng, được mô tả như hình 2.

	MAC	# of days
0	001D20ED4ACA	1983
1	001C55007967	1056
2	001C55007A16	923
3	001C55007A8F	1049
4	001C55007B29	1056
5	001C55007BD7	1007
6	001C55007BFF	1014
7	001C55007C49	986
8	001C55007CF7	1018
9	001C55007D35	1037

Hình 2: Số ngày sử dụng dịch vụ của người dùng sau khi đã xử lý

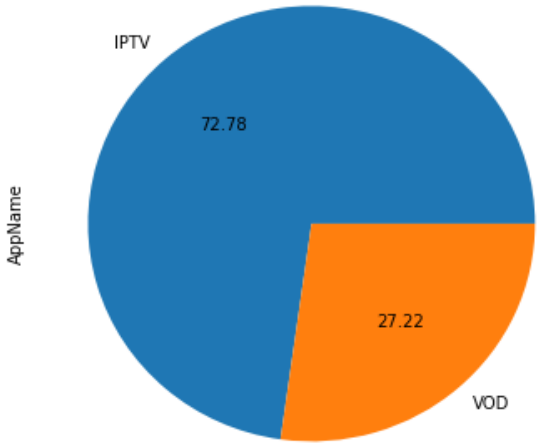
2 Phân tích đặc điểm hành vi người dùng

Các phân tích dưới đây dựa trên tập dữ liệu người dùng đã huỷ hợp đồng. Công cụ sử dụng phân tích: Python, Pivot-Excel. Các bước xử lý khác hỗ trợ cho phân tích:

- Tính tổng `RealTimePlaying` trên cùng một userid thành `TotalTime`
- Loại bỏ tag `FBOX` trong bảng userinfo để phù hợp với bảng dữ liệu datalog. Sau đó merge hai bảng dữ liệu thông qua `MAC`.

Các đặc trưng phân tích hành vi người dùng bao gồm:

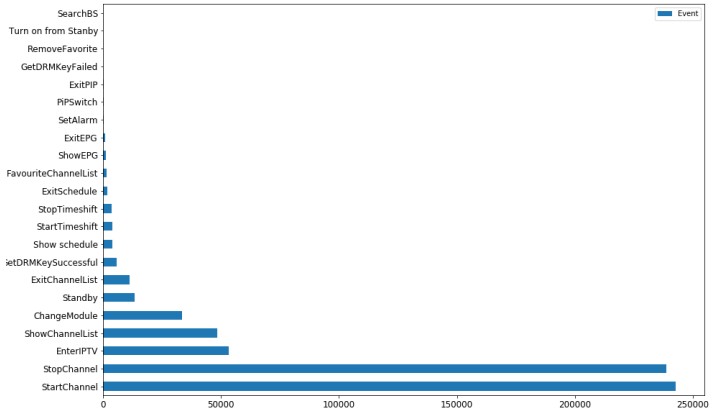
Các ứng dụng thường dùng - AppName. Qua biểu đồ hình , chúng ta có thể thấy rằng đa số người dùng sử dụng IPTV nhiều hơn, gấp 2.7 lần ứng dụng VOD. Từ đó, việc phân tích hành vi người dùng được dựa trên 2 nhóm chính này.



Hình 3: Tỉ lệ các app sử dụng

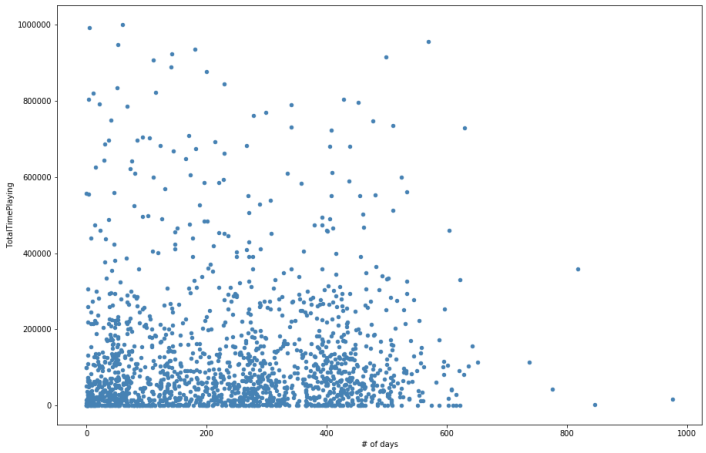
2.1 Nhóm IPTV

Các thao tác người dùng - Event. Người dùng có xu hướng thực hiện thao tác `StartChannel`, và `Stopchannel` chiếm đa số so với các thao tác khác như biểu đồ hình 4. Tuy nhiên, nhóm thao tác được sử dụng nhiều tiếp theo đó là `ChangeModule`, `ShowChannel`, và `EnterIPTV`. Điều này cho thấy rằng, khi số lượng của hai nhóm thao tác này lớn là dấu hiệu cho thấy sự thay đổi nhu cầu sử dụng của dịch vụ của người dùng, và dẫn đến tình trạng huỷ hợp đồng.



Hình 4: Số lượng thao tác người dùng trên IPTV

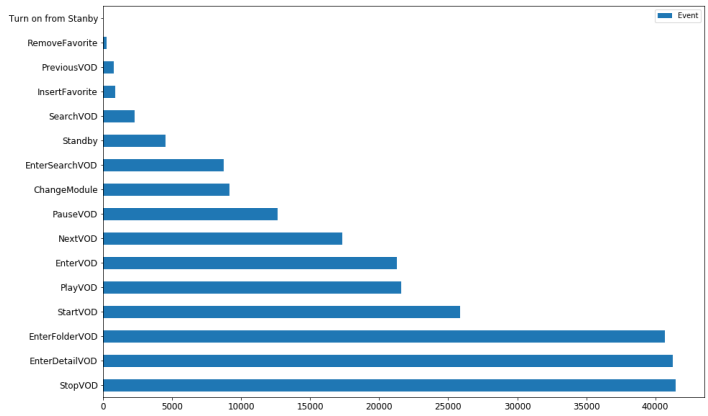
Phân bố giữa Thời gian sử dụng (of days) và Thời lượng xem (RealTime-Playing) Mỗi người dùng có thể xem nhiều lần với thời lượng xem khác nhau. Chính vì vậy, chúng ta sẽ phân tích dựa trên tổng thời lượng xem của từng người dùng (`TotalTimePlaying`). Sau đó, chúng ta kết hợp với dữ liệu thời gian sử dịch vụ để thể hiện phân bố bởi hình 5. Qua đó cho thấy mức độ người dùng sử dụng IPTV đã huỷ hệ thống trong khoảng từ 0-500 (days) với thời lượng xem dưới 200000 (đơn vị thời gian).



Hình 5: Phân bố thời gian sử dụng - Thời gian xem nhóm IPTV

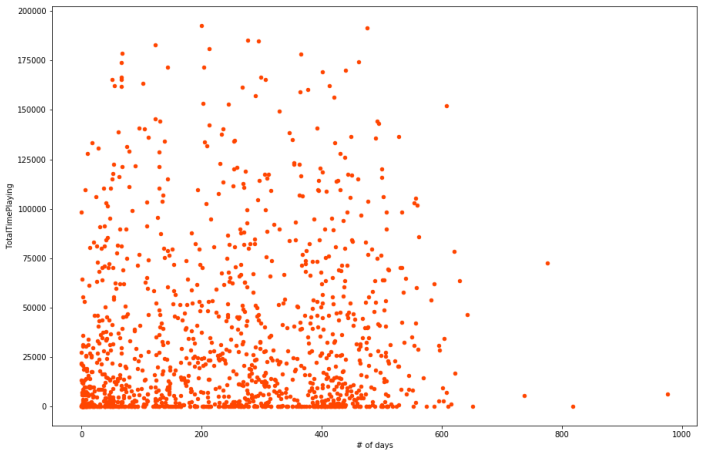
2.2 Nhóm VOD

Từ biểu đồ hình 9, nhóm thao tác gồm `EnterFolderVOD`, `EnterDetailVOD`, và `StopVOD` được thực hiện đa số, ảnh hưởng nhiều đến việc người dùng huỷ hợp đồng.



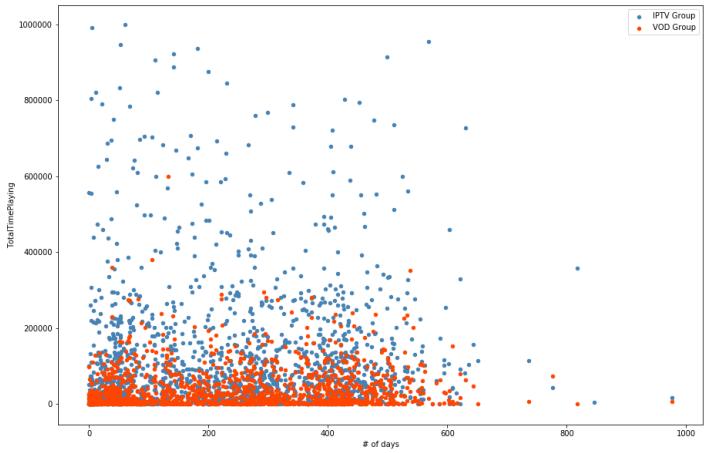
Hình 6: Số lượng thao tác người dùng trên thiết bị

Phân bố giữa Thời gian sử dụng (of days) và Thời lượng xem (RealTime-Playing) Tương tự như nhóm IPTV, phân bố này tập trung nhiều ở mức dưới 500 (days) với thời lượng 25000 (đơn vị thời gian) qua biểu đồ hình 7.



Hình 7: Phân bố thời gian sử dụng - Thời gian xem nhóm VOD

Hình 8 mình hoạ phân bố trên cả 2 nhóm người dùng. Có thể thấy rằng, thời gian sử dụng dịch vụ trên cả 2 ứng dụng dao động từ 0-500 (days). Tuy nhiên, phân bố người dùng sử dụng IPTV có thời lượng xem lớn hơn nhiều so với VOD. Chính vì vậy, việc dự đoán nên cân nhắc trên nhóm VOD khi thời lượng xem thấp.



Hình 8: Phân bố thời gian sử dụng - Thời gian xem trên cả 2 nhóm IPTV - VOD

***Nhận xét:**

- Nhóm người dùng ứng dụng trên IPTV có khả huỷ hợp đồng cao hơn trên nhóm VOD.
- Các thao tác người dùng cũng là yếu tố quan trọng như số lần tắt, mở, chuyển kênh,... lớn cho thấy dấu hiệu của sự nhàm chán dẫn đến khả năng huỷ dịch vụ cao.

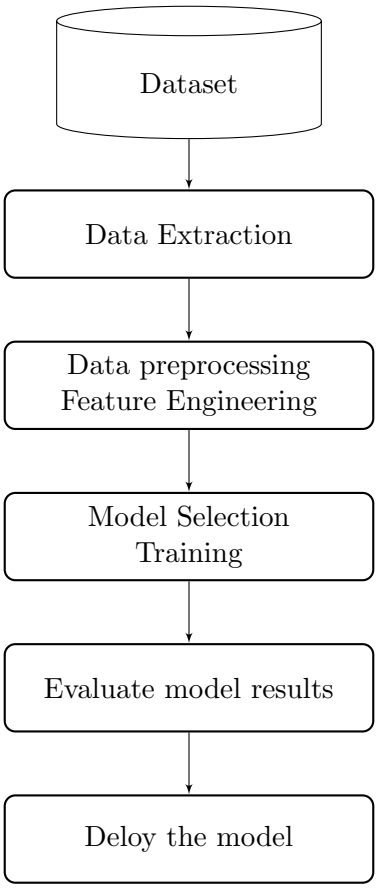
***Đề xuất:** Kết quả dự đoán sẽ hiệu quả hơn khi kết hợp thêm những dữ liệu mô tả sau:

- Dữ liệu thể hiện được tiến trình (process), sự thay đổi từ hành vi này sang hành vi khác trong 1 thời gian ngắn (như 1 ngày) của từng người dùng.
- Dữ liệu liên quan đến thời gian cụ thể của người dùng khi họ sử dụng dịch vụ, như lúc mấy giờ, xem những gì.
- Dữ liệu phân nhóm các loại chương trình (như giải trí, điện ảnh, thể thao,...) của người dùng đã xem ở từng thời điểm.

3 Bài toán "Predicting contract churn/cancellation"

- Bài toán phân loại (Classification)
- Học có giám sát (Suppervised Learning)

Các bước xây dựng hệ thống dự đoán người dùng có khả năng huỷ dịch vụ được mô tả bởi sơ đồ dưới đây:



Hình 9: Sơ đồ mô tả hệ thống

Sau khi xây dựng tập dữ liệu từ các bước rút trích, tiền xử lý như đã trình bày phần 1. Đánh giá các feature bị lỗi, thiếu giá trị để thay thế giá trị phù hợp khác. Lựa chọn phương pháp chuẩn hoá các feature như:

- Input Variables:
  - `MAC`
  - `AppName (IPTV, VOD)`: Feature binarization
  - `Event`, `ItemID`: OneHotEncoder
  - `RealTimePlaying`, `OfDays`: Scaling features to a range
- Output Variables: Biến dự đoán kết quả dưới dạng nhị phân (1 = có khả năng huỷ hợp đồng, và 0 ngược lại).

Model Selection and Training.

- Chia tập dữ liệu ngẫu nhiên với 70% cho quá trình huấn luyện, 30% cho quá trình kiểm tra.
- Thuật toán: Logistic Regression, Decision Tree, hoặc Random Forest

Evaluate model results.

- Các hệ số đánh giá: Accurrancy, Error Rate, Recall, Precision