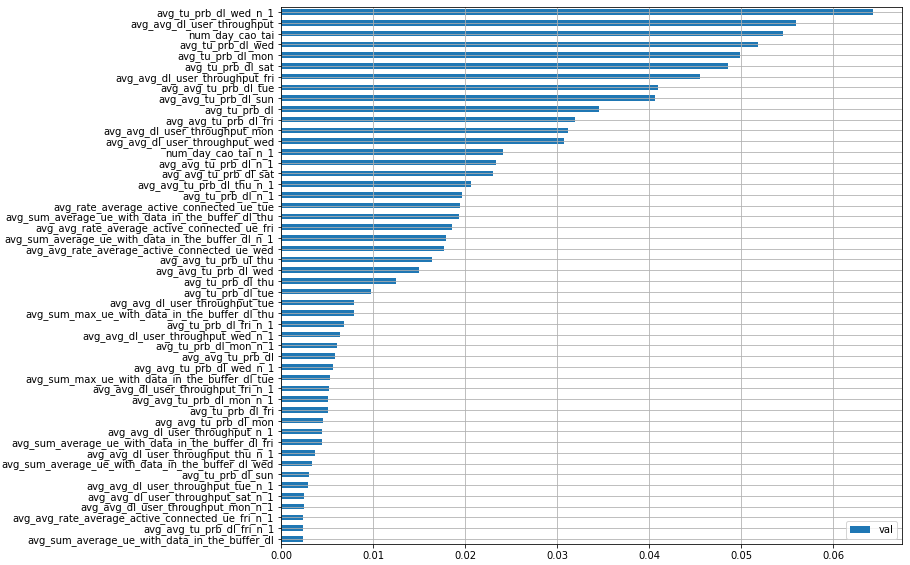
1. **Cách làm hiện tại**

* ***Cách đánh nhãn***: Dựa trên các điều kiện cảnh báo nâng cấp từ 2 trường dữ liệu TU\_PRB\_DL và USER\_DL\_THROUGHPUT xác định sector có cao tải hay không (nhãn 1 và 0 trong bài toán phân lớp) trong thời điểm 3 tháng tiếp theo.
* ***Cách chia tập train/valid/test***: Chia tập dựa trên các khoảng thời gian khác nhau tùy vào mục đích đánh giá và dự đoán của mô hình. Ví dụ:
  + Train: sử dụng toàn bộ dữ liệu từ 04/2019 đến 06/2021, trong đó sử dụng dữ liệu 04/2019 và nhãn tại 07/2019, sử dụng dữ liệu 05/2019 và nhãn tại 08/2019, sử dụng dữ liệu 06/2019 và nhãn tại 09/2019, …
  + Valid: sử dụng dữ liệu từ 07/2021 đến 09/2021
  + Test: sử dụng dữ liệu từ 10/2021 đến 12/2021

1. **Vấn đề gặp phải**

* ***Vấn đề:*** Hiện tại phân bố dữ liệu của các top feature đặc trưng có điểm quan trọng (importance score) trên các tập train, valid, test khác nhau.
* ***Cơ sở lý thuyết đánh giá***: Theo Andrew Ng, phân bố dữ liệu trên các tập cần có phân bố giống nhau, và cần phản ánh dữ liệu có thể có tại thời điểm dự đoán trong tương lai (Andrew Ng, Machine Learning Yearning)
* ***Phương pháp đánh giá***: Trên từng mô hình áp dụng với mỗi tỉnh, lấy các top feature và đánh giá phân bố dữ liệu của đặc trưng trên từng nhãn với các tập train/test. Áp dụng kiểm định thống kê để xác định phân bố dữ liệu có tương đồng nhau không (sử dụng kiểm định Welch’s t-test để kiểm định sự tương đồng giữa giá trị trung bình của hai phân phối trên tập train và test.
* ***Ví dụ trên Hà Nội*** :
  + Bước 1: Dựa trên các top features, đánh giá trên các features đại diện có giá trị importance score cao nhất.



Feature lựa chọn để đánh giá:

['avg\_sum\_average\_ue\_with\_data\_in\_the\_buffer\_dl',

'avg\_avg\_dl\_user\_throughput',

'avg\_avg\_rate\_average\_active\_connected\_ue', 'avg\_sum\_max\_ue\_with\_data\_in\_the\_buffer\_dl',

'avg\_rate\_average\_active\_connected\_ue',

'avg\_avg\_tu\_prb\_dl',

'avg\_avg\_tu\_prb\_ul',

'avg\_tu\_prb\_dl',

'num\_day\_cao\_tai']

* + Bước 2: Đánh giá phân bố dữ liệu của các features trên tập train và test:

Ghi chú ký hiệu:

train\_1: tập train nhãn 1; train\_0: tập train nhãn 0

test\_20211001\_1: tập test tháng 10/2021 nhãn 1

Giá trị trong mỗi ô thể hiện kết quả của kiểm định t-test (1: cùng phân bố, 0: khác phân bố)

Để đảm bảo tính ứng dụng của mô hình thì tập train và test cần có phân bố giống nhau cả nhãn 0 và nhãn 1. Trong bảng dưới đây, để tập train và tập test có cùng phân bố trên một feature thì từng feature cần có giá trị trong hai cột cùng màu phải có cùng giá trị và bằng 1.

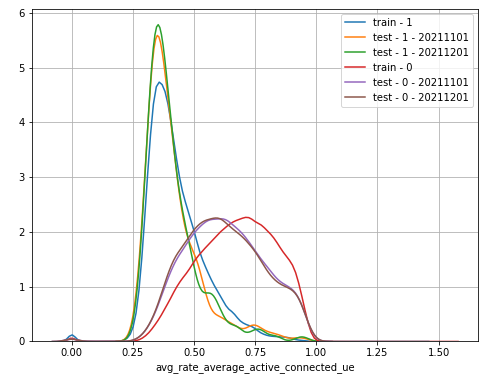
Bảng dưới chỉ ra không có feature nào có cùng phân bố giữa tập train và test:

| **features** | **train\_1 vs**  **test\_20211001\_1** | **train\_0 vs**  **test\_20211001\_0** | **train\_1 vs**  **test\_20211101\_1** | **train\_0 vs**  **test\_20211101\_0** | **train\_1 vs**  **test\_20211201\_1** | **train\_0 vs**  **test\_20211201\_0** |
| --- | --- | --- | --- | --- | --- | --- |
| avg\_sum\_average\_ue\_with\_data\_in\_the\_buffer\_dl | 1 | 0 | 1 | 0 | 1 | 0 |
| avg\_avg\_dl\_user\_throughput | 0 | 0 | 0 | 0 | 0 | 1 |
| avg\_avg\_rate\_average\_active\_connected\_ue | 0 | 0 | 0 | 0 | 0 | 0 |
| avg\_sum\_max\_ue\_with\_data\_in\_the\_buffer\_dl | 1 | 0 | 1 | 0 | 1 | 0 |
| avg\_rate\_average\_active\_connected\_ue | 0 | 0 | 0 | 0 | 0 | 0 |
| avg\_avg\_tu\_prb\_dl | 0 | 0 | 0 | 0 | 0 | 0 |
| avg\_avg\_tu\_prb\_ul | 0 | 0 | 0 | 0 | 0 | 0 |
| avg\_tu\_prb\_dl | 0 | 0 | 0 | 0 | 0 | 0 |
| num\_day\_cao\_tai | 0 | 0 | 0 | 0 | 0 | 0 |

Ví dụ phân phối của biến avg\_rate\_average\_active\_connected\_ue:

So sánh phân phối nhãn 1 giữa các đường: train - 1 vs test - 1 - 20211101; train - 1 vs test - 1 - 20211201 → Kiểm định t-test cho kết quả không cùng phân bố.

So sánh phân phối nhãn 0 giữa các đường: train - 0 vs test - 0 - 20211101; train - 0 vs test - 0 - 20211201 → Kiểm định t-test cho kết quả không cùng phân bố.



Đánh giá: phân phối nhãn 1/ nhãn 0 của tập train và tập test không ổn định qua các tháng.

1. **Tính khả thi phương án thay thế**

* ***Phương án thay thế:*** Tìm trong các tháng của tập train để lựa chọn ra các tháng có cùng phân bố dữ liệu với tháng của tập test. Cùng phân bố dữ liệu được tính trên sự tương đồng của phân bố dữ liệu của nhãn 0/1 của top các features trên tập train và test.
* ***Kết quả thử nghiệm***:
  + Lựa chọn các tháng trong tập train (từ 04/2019 - 06/2021) để lựa chọn ra các tháng có cùng phân bố dữ liệu với mỗi tháng của ba tập test khác nhau (tháng 10, 11, 12/2021)
  + Thử nghiệm trên điều kiện chỉ cần 1 feature duy nhất để tìm các tháng có phân bố dữ liệu tương đồng. Kết quả tìm kiếm trả ra như sau:
    - Tháng 10/2021: không tìm được tháng nào cùng phân bố dữ liệu
    - Tháng 11/2021: tìm được tháng 3/2021 cùng phân bố dữ liệu
    - Tháng 12/2021: tìm được tháng 3/2021 cùng phân bố dữ liệu

=> Nhận xét: Với dữ liệu hiện tại, việc áp dụng phương án thay thế là không khả thi do với điều kiện trên chỉ có một tháng dùng để huấn luyện trên một feature có tính tương đồng, mô hình không có khả năng phân loại.

1. **Đánh giá/Kết luận**

Với các phân tích và đánh giá như trên, nhóm đề xuất như sau:

* Tại thời điểm hiện tại, với bộ feature dựa trên các trường dữ liệu đang có, việc áp dụng mô hình phân lớp để dự báo sector có cao tải trong thời gian tiếp theo không là không khả thi do phân bố dữ liệu của các feature không giống nhau trong các tháng dùng để huấn luyện và dự đoán. Do đó, hướng tiếp cận theo phân lớp sẽ tạm thời pending.
* Tiếp tục thực hiện khảo sát mô hình dữ liệu để bổ sung thêm các feature mới có tính tương đồng và có khả năng dự đoán.
* Tiếp tục phối hợp với OS tiếp cận theo hướng sử dụng các mô hình timeseries trên các chỉ số mục tiêu, hạn chế việc ảnh hưởng của các feature đang áp dụng trong bài toán phân lớp.