Dự đoán số ngày nằm viện của các ca bệnh nhiễm SARS-CoV-2 tại Việt Nam

Ngô Hoàng Anh (Ecole Polytechnique - Pháp)
Hoàng Thái Nam (Beloit College - Mỹ)
Nguyễn Tuấn Khôi (University of Melbourne - Úc)

Mục đích

Dự đoán thời gian nằm viện của các ca nhiễm SARS-CoV-2 tại Việt Nam nhằm chủ động kiểm soát dịch bệnh và phân bố nguồn lực

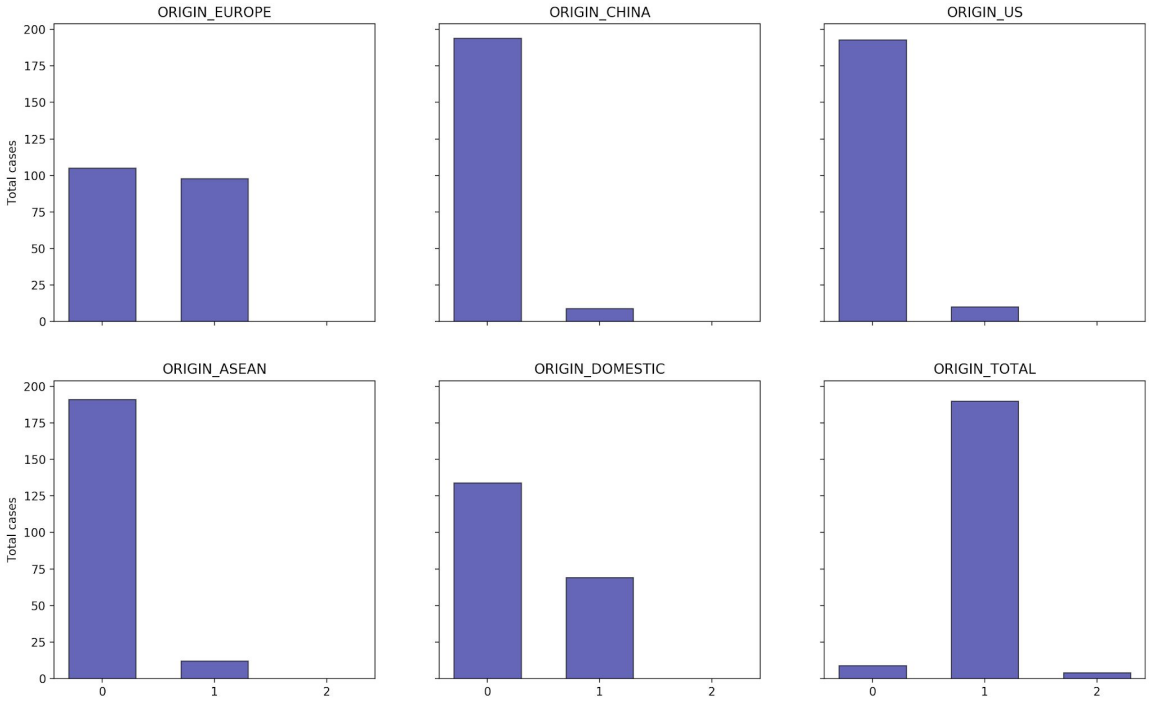
Xây dựng dữ liệu

- Sử liệu nguồn dữ liệu thô (được cung cấp bởi TS Nguyễn Thu Anh)
- Lọc và chuyển dữ liệu thành dạng file .csv nhờ vào file .exe (viết bằng Java)
- Tinh chỉnh này dữ liệu dưới dạng biến số, gồm 16 biến, 2 dạng biến (binary và số)

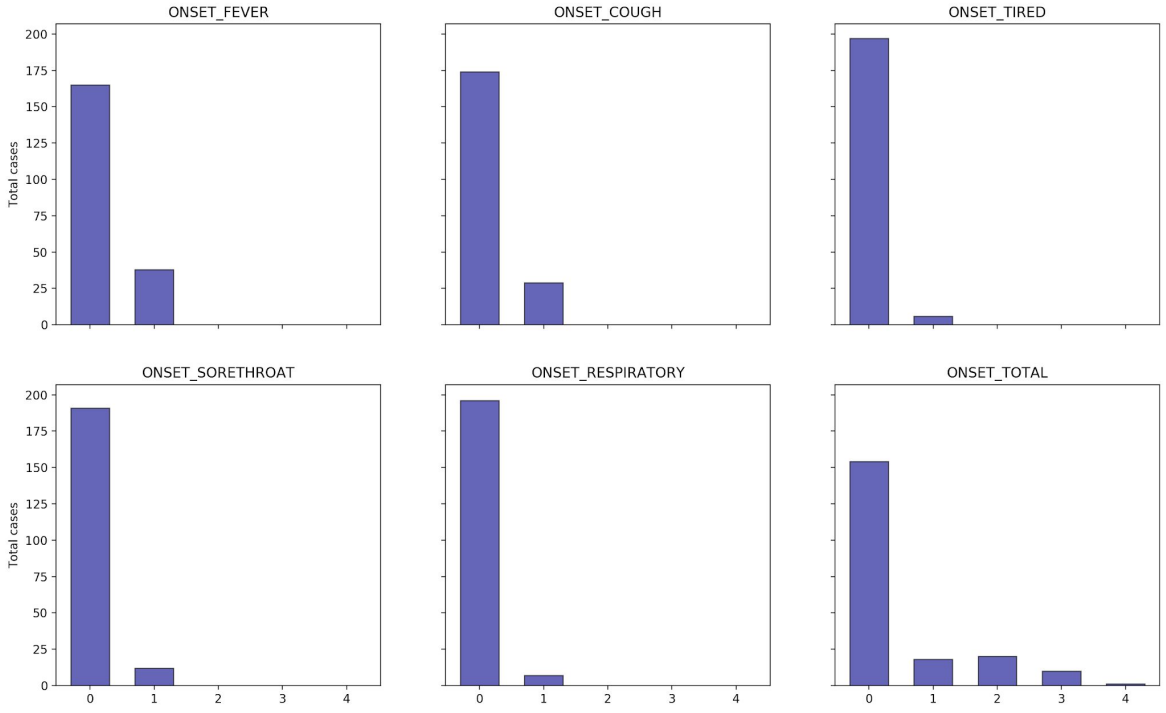
Xây dựng dữ liệu

PATIENT_NO	ONSET_FEVER
AGE	ONSET_COUGH
SEX	ONSET_TIRED
ORIGIN_EUROPE	ONSET_SORETHROAT
ORIGIN_CHINA	ONSET_RESPIRATORY
ORIGIN_US	BACKGROUND DISEASE
ORIGIN_ASEAN	HOSPITAL
ORIGIN_DOMESTIC	RISK_SCORE
DAYS_OF_ONSET	LOS

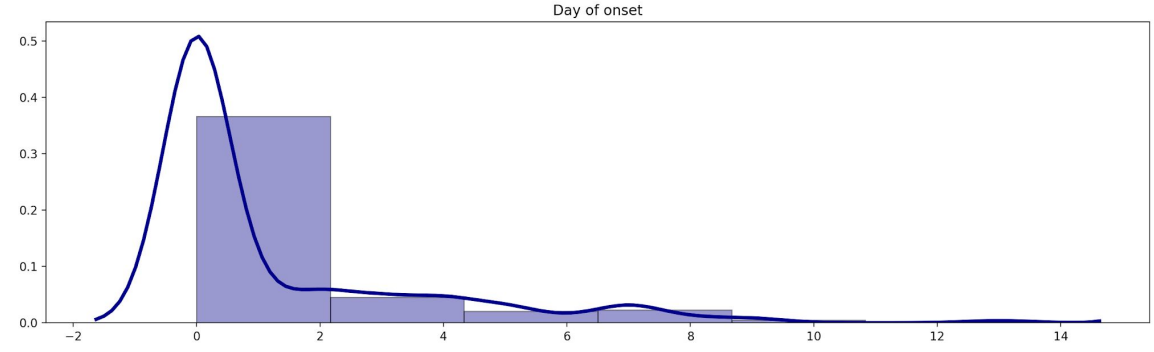
Minh họa

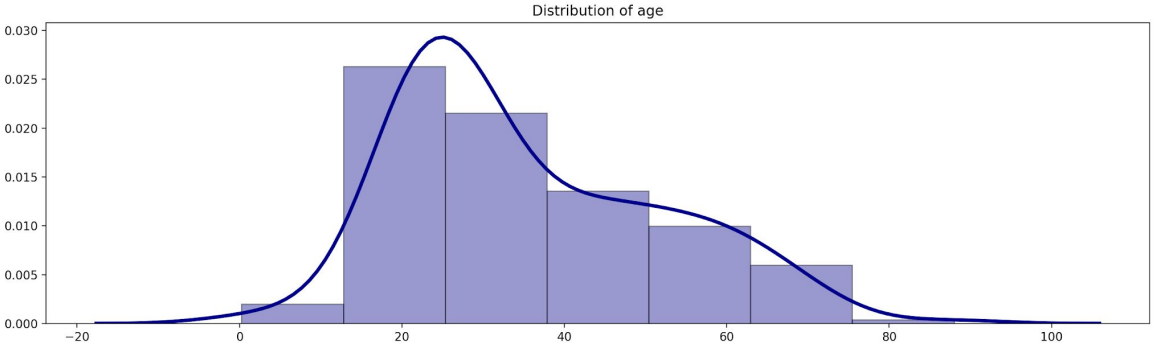

ORIGIN_EUROPE
ORIGIN_CHINA
ORIGIN_US
ORIGIN_ASEAN
ORIGIN_DOMESTIC
ORIGIN_TOTAL
Total cases
Total cases
0
25
50
75
100
125
150
175
200
0
1
2

Minh họa


ONSET_FEVER
ONSET_COUGH
ONSET_TIRED
ONSET_SORETHROAT
ONSET_RESPIRATORY
ONSET_TOTAL
Total cases
Total cases
0
25
50
75
100
125
150
175
200
0
1
2
3
4

Minh họa

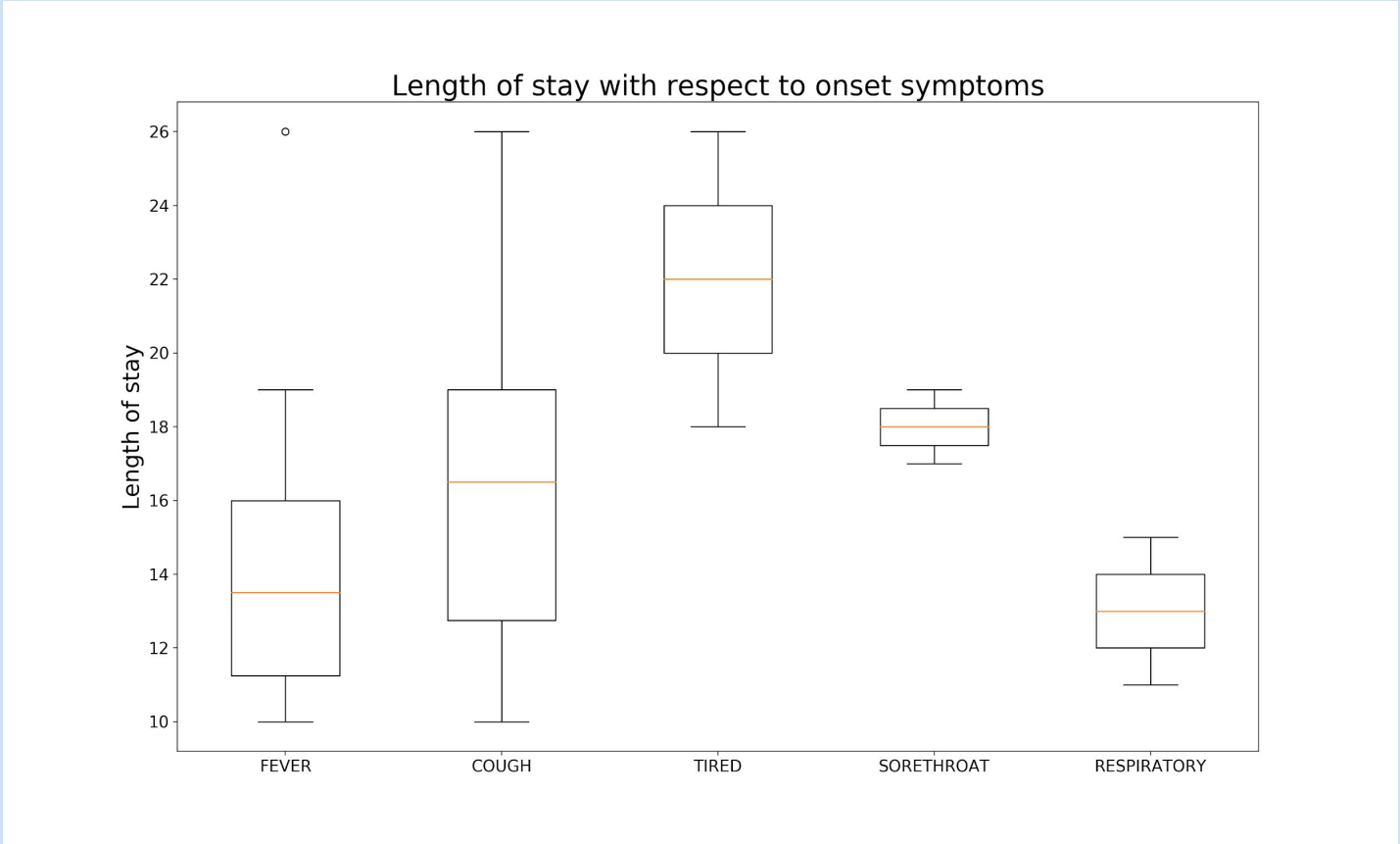

Day of onset
0.5
0.4
0.3
0.2
0.1
0.0
−2
0
2
4
6
8
10
12
14


Distribution of age
0.030
0.025
0.020
0.015
0.010
0.005
0.000
−20
0
20
40
60
80
100

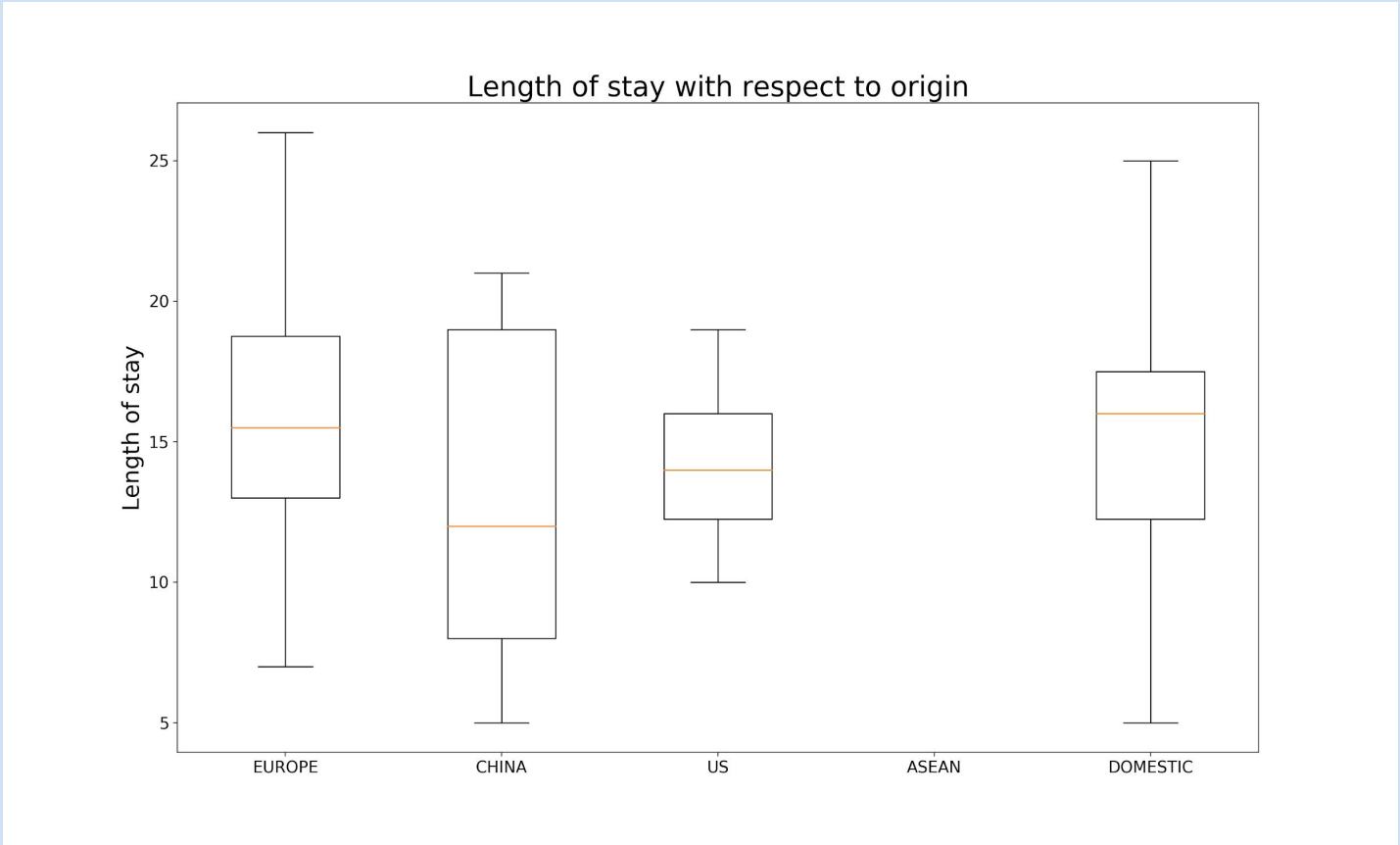
Minh họa


Length of stay with respect to gender
Length of stay
25
20
15
10
5
Male
Female

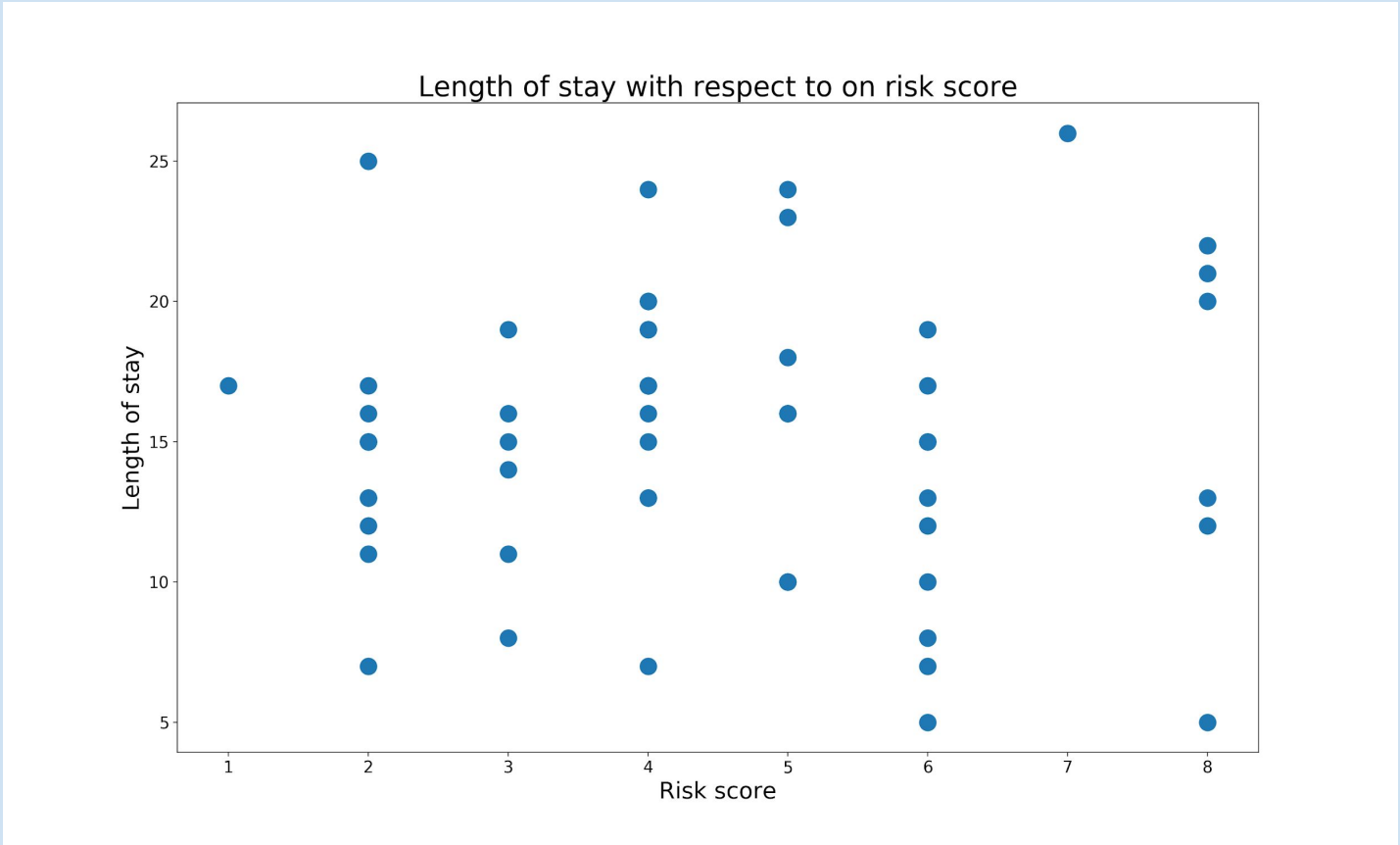
Minh họa


Length of stay with respect to onset symptoms
Length of stay
26
24
22
20
18
16
14
12
10
FEVER
COUGH
TIRED
SORETHROAT
RESPIRATORY

Minh họa


Length of stay with respect to origin
Length of stay
25
20
15
10
5
EUROPE
CHINA
US
ASEAN
DOMESTIC

Minh họa


Length of stay with respect to on risk score
Length of stay
Risk score
25
20
15
10
5
1
2
3
4
5
6
7
8

Các mô hình được sử dụng

- Linear model:
  - Linear Regression
  - Ridge Regression
  - Lasso Regression
- Tree model:
  - Decision Tree
  - Random Forest
- Gradient Descent
  - Stochastic Gradient Descent (SGD)

Các mô hình được sử dụng

- Tree model:
  - Decision Tree
  - Random Forest
- Boosting Machine
  - Gradient Boosting Machine (GBM)
  - Extreme Gradient Boosting (XGB)
- k-Nearest Neighbors (kNN)
- Support Vector Machine (SVM)

Parameter Tuning

- Các mô hình có quy trình hoạt động khác nhau, đòi hỏi thiết lập khác nhau => vận hành và đưa ra dự đoán chuẩn xác tối ưu.
- Tồn tại các phương pháp có thể thay đổi tham số cho phù hợp với mục đích làm tăng độ chuẩn xác.
- Cần căn chỉnh và thử nghiệm các siêu tham số (hyperparameter) khác nhau, chọn ra siêu tham số mang lại độ chính xác cao nhất.

Cross validation

- Chia mô hình thành các phần dữ liệu training set và testing set

- K-fold cross validation: khắc phục phương pháp kiểm chứng chéo thông thường bằng cách chia dữ liệu thành *k* phần bằng nhau

Feature importance

- Để có thể tối ưu hóa tốc độ làm việc và tránh phức tạp hóa mô hình, cần chọn lọc biến số có ảnh hưởng cao.

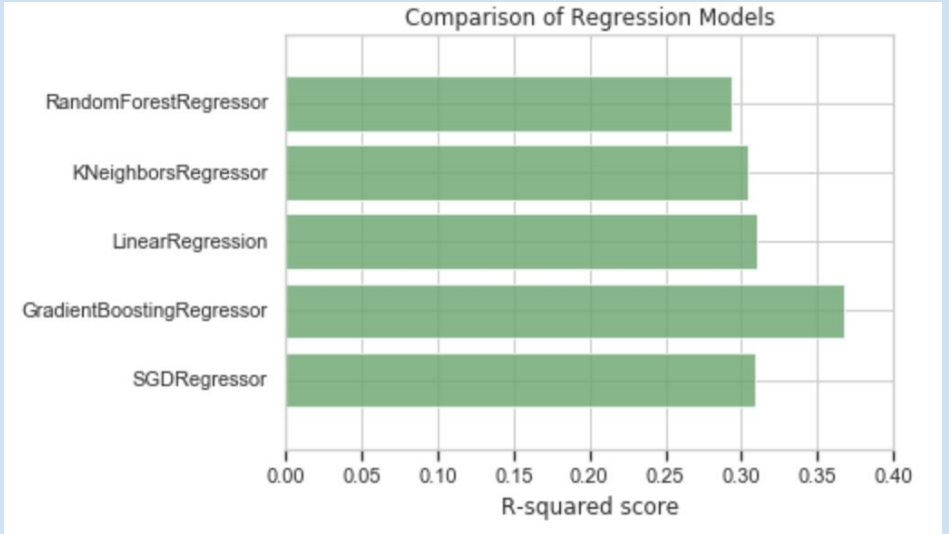
- Yếu tố quyết định mức độ ảnh hưởng của một biến số được gọi là tầm quan trọng - feature importance.

Các công trình đã công bố

- Những báo cáo được công bố trước đây, nhằm mục đích dự đoán thời gian nằm bệnh của các bệnh nhân thuộc tất cả các khoa, phòng bao gồm:
  - Predicting Length-of-Stay at Hospitals (Project Notebook), Daniel Cummings (Deep learning Scientist at intel)
  - Predicting inpatient flow at a major hospital using interpretable analytics (Dimitris Bertsimas, Jean Pauphilet, Jennifer Stevens, Manu Tandon)

Hiệu quả các mô hình đã được sử dụng

	LR	CART	OT	RF	GBT
Classification: remaining length of stay < 1 day					
AUC	0.826	0.807	0.810	**0.843**	0.839
MAE in # daily discharges, no.	8.6	**6.0**	6.4	6.2	7.8
MRE in # daily discharges, %	8.7	6.0	6.5	**5.8**	7.6
Out-of-sample R^2	0.730	0.868	0.847	0.841	0.804
Classification: remaining length of stay < 2 days					
AUC	0.809	0.786	0.790	0.815	**0.822**
Classification: overall length of stay < 7 days					
AUC	0.818	0.775	0.776	0.813	**0.820**
AUC at day 1	0.827	0.795	0.797	0.828	**0.830**
AUC at day 2	**0.807**	0.752	0.752	0.800	0.804
Classification: overall length of stay < 14 days					
AUC	**0.826**	0.777	0.777	0.820	0.794



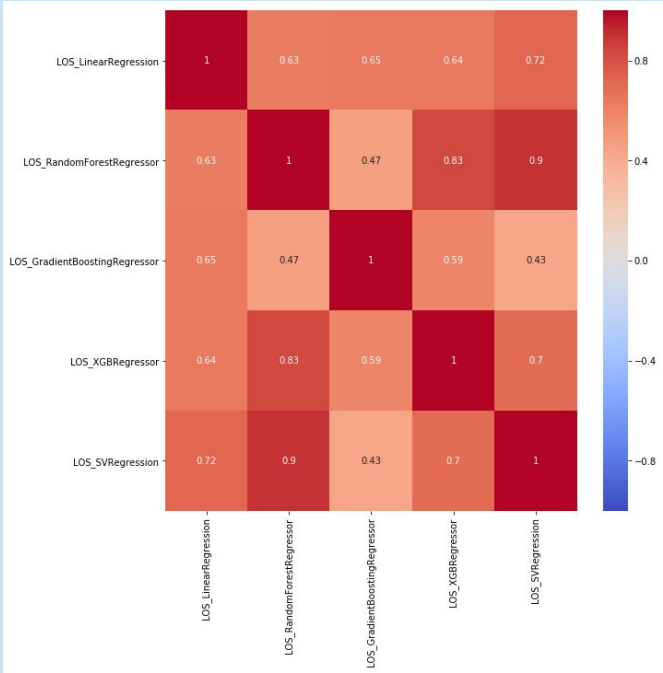
Các công trình đã công bố

- Những báo cáo được công bố trước đây, nhằm mục đích dự đoán thời gian nằm bệnh của các bệnh nhân thuộc tất cả các khoa, phòng bao gồm:
  - Predicting Length-of-Stay at Hospitals (Project Notebook), Daniel Cummings (Deep learning Scientist at intel)
  - Predicting inpatient flow at a major hospital using interpretable analytics (Dimitris Bertsimas, Jean Pauphilet, Jennifer Stevens, Manu Tandon)

Lựa chọn mô hình phù hợp

- Dựa trên các mô hình đã chạy, cùng với kết quả cross-validation và phân loại mô hình, có 5 trên tổng số 9 mô hình được chọn như sau:
  - Linear Regression (mô hình benchmark)
  - Random Forest
  - Gradient Boosting
  - Extreme Gradient Boosting
  - Support Vector Machine

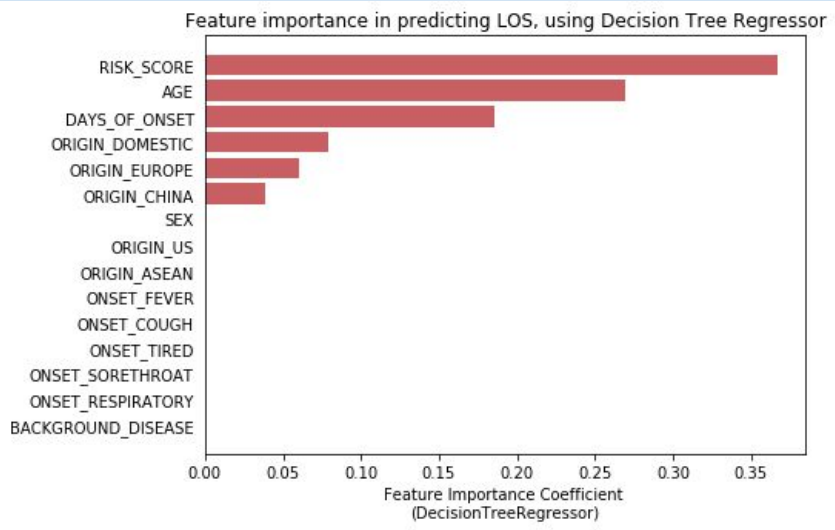
Lựa chọn mô hình phù hợp

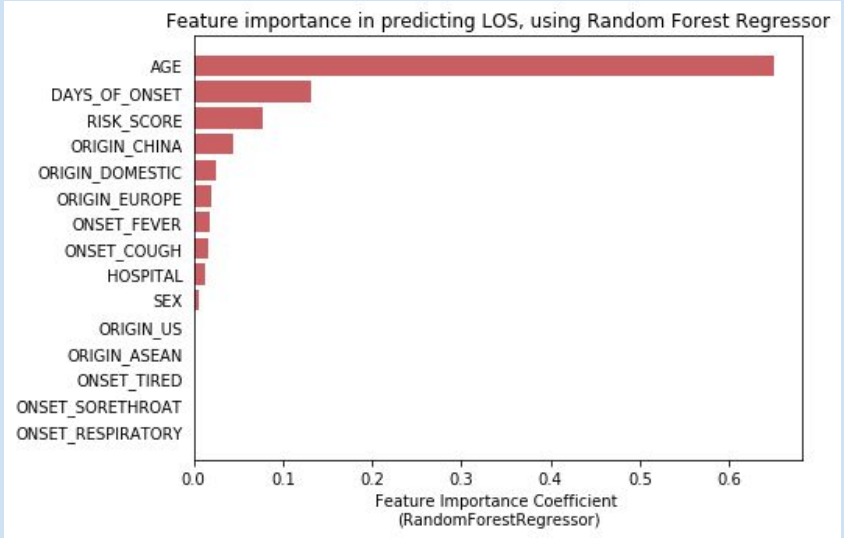


Lựa chọn mô hình phù hợp

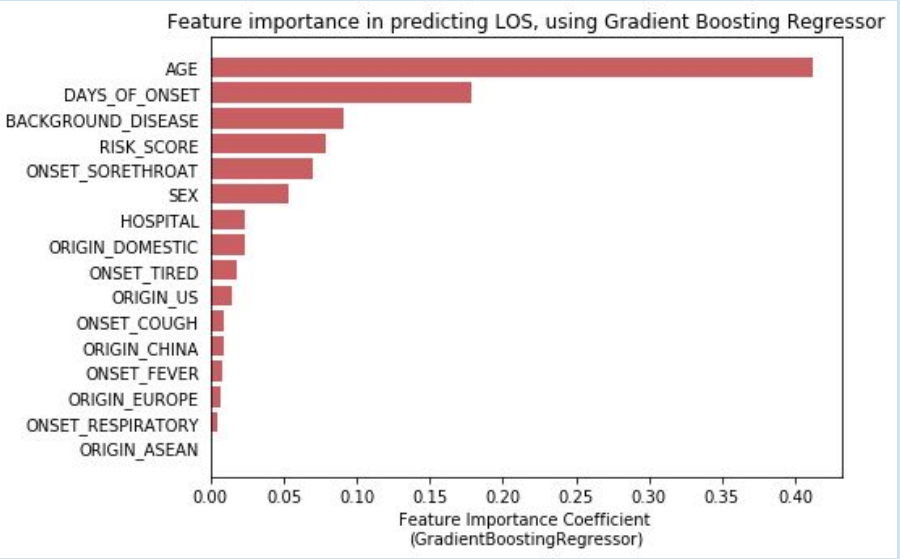
- Sau khi xem xét thêm, có hai mô hình được xem là hiệu quả nhất, mang lại tác dụng chẩn đoán thực tiễn nhất:
  - Gradient Boosting Regressor: Đã được chính minh là mô hình hoạt động hiệu quả nhất trong cả hai công bố được nhắc đến trước đó
  - Extreme Gradient Boosting Regressor: Bản nâng cấp của Gradient Boosting, được sử dụng rất nhiều trong các bài toán dự đoán gần đây.
  - SVR: Support Vector Regressor, dựa trên SVM là một classification model rất mạnh trong machine learning

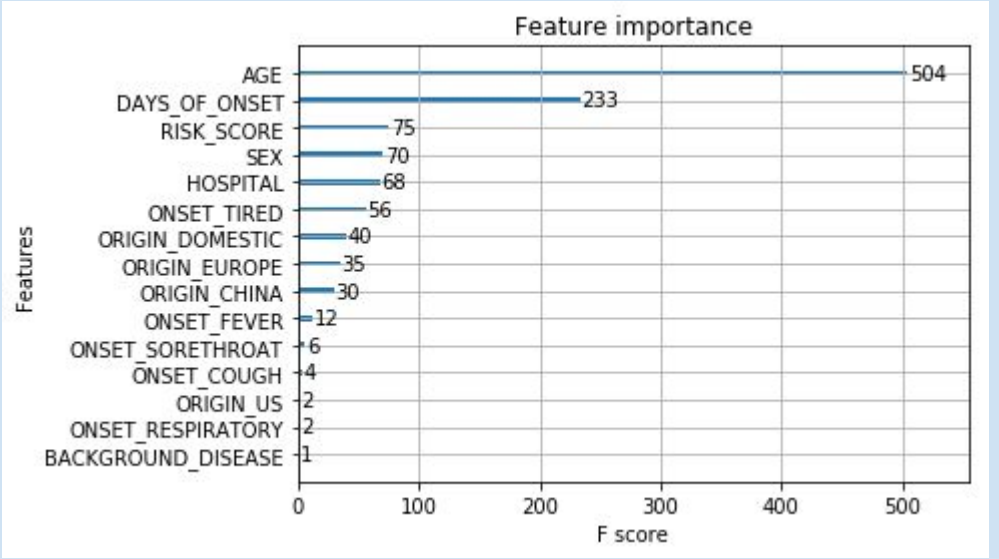
Feature Importance





Feature Importance





Hướng đi tiếp theo

- Model tuning: Có rất nhiều hyperparameter khác nhau có thể ảnh hưởng tới hiệu quả của model, đặc biệt là khi dữ liệu thay đổi liên tục
- Các biến số dịch tễ + y tế thêm vào (ví dụ như với dữ liệu tại Mỹ, có các thông tin y tế như BMI, số lần vào bệnh viện trong 6 tháng vừa rồi, etc.)
- Xây dựng phần mềm để ứng dụng tại các bệnh viện để có khả năng dự đoán và phân bổ nguồn lực ngay khi tiếp nhận bệnh nhân mới