Heritage Health Prize Report

Tuan Nguyen

December 19, 2020

Chapter 1

Data Preprocessing

1.1 Strategies

Dataset được cung cấp chứa dữ liệu claim của bệnh nhân trong 3 năm (Y1, Y2 và Y3). Mục tiêu của bài toán là dùng dữ liệu của 3 năm này để dự đoán số ngày nằm viện của bệnh nhân đó trong năm thứ 4 (Y4).

Tuy nhiên, chúng ta không thể có dữ liệu của Y4 để xây dựng mô hình từ 3 năm đầu tiên do chính sách của Kaggle. Vì vậy, trong giới hạn của project này, chúng ta sẽ dùng dữ liệu Y1 và Y2 để làm train set và dùng Y3 để làm test set.

Dữ liệu hiện có chứa thông tin các lần claim trong năm của một bệnh nhân. Để có thể dự đoán số ngày nằm viện cho một bệnh nhân trong năm kế tiếp, dữ liệu claim cần được xử lý phù hợp để chuyển tất cả thông tin claim thành thông tin tổng hợp của một bệnh nhân.

Khi quan sát tập dữ liệu, ta nhận thấy có một số dữ liệu cố định của bệnh nhân, không phụ thuộc vào năm, ví dụ: "Sex", "AgeAtFirstClaim". Các thông tin này không thay đổi dù bệnh nhân có claim bao nhiêu lần đi chăng nữa. Cùng với đó là một số dữ liệu thay đổi theo mỗi lần claim, chẳng hạn: "PayDelay", "PrimaryConditionGroup", ...

Do đó, sẽ có 2 chiến lược xây dựng dữ liệu cho mô hình:

Strategy A: Chỉ dùng dữ liệu của 1 năm để dự đoán cho năm sau. Các thông tin trong bảng Claims sẽ được tổng hợp theo từng năm.

Strategy B: Kết hợp dữ liệu của tất cả các năm trong quá khứ để dự đoán cho năm tiếp theo. Dữ liệu claim của Y1 và Y2 trong bảng Claims sẽ được kết hợp lại.

Trong project này, ta sẽ xây dựng cả 2 mô hình để đánh giá chiến lược nào sẽ cho kết quả tốt hơn.

1.2 Data Manipulation

Phần này sẽ mô tả cách thức xử lý từng feature để tạo ra dataset cho 2 chiến lược ở trên.

1.2.1 Bảng Members

Bảng này chứa dữ liệu chung của bệnh nhân và không phụ thuộc thời gian.

MemberID: Đã được de-identification, feature này chỉ được dùng làm index cho các feature khác và không dùng trong quá trình xây dựng mô hình.

AgeAtFirstClaim: Biểu thị giá trị rời rạc cho khoảng độ tuổi và sẽ được chuyển thành dữ liệu số là giá trị trung bình của khoảng đô tuổi đó. Xem bảng sau:

Giá trị ban đầu	Giá trị thay thế
0-9	5
10-19	15
20-29	25
30-39	35
40-49	45
50-59	55
60-69	65
70-79	75
80+	85
NaN	-1

Table 1.1: AgeAtFirstClaim

Sex: Giới tính của bệnh nhân sẽ được chuyển thành one-hot vector. Gồm 3 giá trị: M, F và NaN

1.2.2 Bảng Claims

Chứa dữ liệu claim của bệnh nhân, mỗi feature trong bảng này sẽ được xử lý theo Y1, Y2 và kết hợp giữa Y1 và Y2 (Y12).

ClaimCount: Đếm số lượng claim của mỗi bệnh nhân theo Y1, Y2 và Y12.

Provider: Đếm số lượng Provider duy nhất của từng bệnh nhân theo Y1, Y2 và Y12.

Vendor: Đếm số lượng Vendor duy nhất của từng bệnh nhân theo Y1, Y2 và Y12.

PCP: Đếm số lượng PCP duy nhất của từng bệnh nhân theo Y1, Y2 và Y12.

Specialty: Đếm số lượng mỗi giá trị của Speciaty cho từng bệnh nhân theo Y1, Y2 và Y12.

PlaceSvc: Đếm số lượng mỗi giá trị của PlaceSvc cho từng bệnh nhân theo Y1, Y2 và Y12.

PayDelay: Dữ liệu sẽ được chuyển thành kiểu số. Giá trị 162+ chuyển thành 162. Sau đó tính toán các giá trị thống kê như min, max, avg, std, sum của từng bệnh nhân theo Y1, Y2 và Y12.

LengthOfStay: Dữ liệu được chuyển thành trung bình số ngày của mỗi giá trị và tính toán các giá trị thống kê như min, max, avg, std, sum của từng bệnh nhân theo Y1, Y2 và Y12. Xem bảng sau:

Giá trị ban đầu	Giá trị thay thế
1 day	1
2 days	2
3 days	3
4 days	4
5 days	5
6 days	6
1- 2 weeks	11
2- 4 weeks	21
4- 8 weeks	42
8- 12 weeks	84
12- 26 weeks	133
26+ weeks	182

Table 1.2: LenghOfStay

DSFS: Dữ liệu được chuyển thành số tháng và tính toán các giá trị thống kê min, max của từng bệnh nhân theo Y1, Y2 và Y12. Xem bảng sau:

CharlsonIndex: Dữ liệu được chuyển thành số và tính toán các giá trị thống kê min, max, avg của từng bệnh nhân theo Y1, Y2 và Y12. Xem bảng sau:

PrimaryConditionGroup: Đếm số lượng mỗi giá trị của PrimaryConditionGroup cho từng bệnh nhân theo Y1, Y2 và Y12.

ProcedureGroup: Đếm số lượng mỗi giá trị của ProcedureGroup cho từng bệnh nhân theo Y1, Y2 và Y12.

Giá trị ban đầu	Giá trị thay thế
0- 1 month	1
1- 2 months	2
2- 3 months	3
3- 4 months	4
4- 5 months	5
5- 6 months	6
6- 7 months	7
7- 8 months	8
8- 9 months	9
9-10 months	10
10-11 months	11
11-12 months	12

Table 1.3: DSFS

Giá trị ban đầu	Giá trị thay thế
0	0
1-2	2
3-4	4
5+	6

Table 1.4: CharlsonIndex

1.2.3 Bảng LabCount

LabCount: Chuyển giá trị 10+ thành 10 và tính toán các giá trị thống kê như min, max, avg, std, sum và số lượng claim (LabClaimCount) của từng bệnh nhân theo Y1, Y2 và Y12.

1.2.4 Bảng DrugCount

DrugCount: Chuyển giá trị 7+ thành 7 và tính toán các giá trị thống kê như min, max, avg, std, sum và số lượng claim (LabClaimCount) của từng bệnh nhân theo Y1, Y2 và Y12.

1.3 Dataset building

Dữ liệu ban đầu sau khi tiền xử lý sẽ được dùng để xây dựng dữ liệu đầu vào theo 2 chiến lược bên dưới.

1.3.1 Strategy A

Ở chiến lược này, ta chỉ sử dụng dữ liệu một năm để xây dựng mô hình dự đoán kết quả của năm tiếp theo. Vì vậy tập dữ liệu được phân chia như bảng bên dưới:

Training set	Testing set
Claims Y1 Outcome: DaysInHospital Y2	Claims Y2 Outcome: DaysInHospital Y3

Table 1.5: Training set và testing set cho Strategy A

Các feature trong Strategy A:

Feature	Description
MemberID	Chỉ dùng làm index
AgeAtFirstClaim	Tuổi lúc claim lần đầu
Sex	Giới tính
ClaimCount	Số lượng claim trong năm
ProviderCount	Số lượng Provider duy nhất trong năm
VendorCount	Số lượng Vendor duy nhất trong năm
PCPCount	Số lượng PCP duy nhất trong năm
SpecialtyCount_<value>	Số lượng claim từng giá trị trong năm
PlaceSvcCount_<value>	Số lượng claim từng giá trị trong năm
PayDelayMin	Min PayDelay trong năm
PayDelayMax	Max PayDelay trong năm
PayDelayAvg	Avg PayDelay trong năm
PayDelayStd	Std PayDelay trong năm
PayDelaySum	Sum PayDelay trong năm
LengthOfStayMin	Min LengthOfStay trong năm
LengthOfStayMax	Max LengthOfStay trong năm
LengthOfStayAvg	Avg LengthOfStay trong năm
LengthOfStayStd	Std LengthOfStay trong năm
LengthOfStaySum	Sum LengthOfStay trong năm
LengthOfStayCountNan	Số lượng LengthOfStay là NaN trong năm
DSFSMin	Min DSFS trong năm
DSFSMax	Max DSFS trong năm
CharlsonIndexMin	Min CharlsonIndex trong năm
CharlsonIndexMax	Max CharlsonIndex trong năm
CharlsonIndexAvg	Avg CharlsonIndex trong năm
PrimaryConditionGroupCount_<value>	Số lượng claim từng giá trị trong năm
ProcedureGroupCount_<value>	Số lượng claim từng giá trị trong năm

LabCountMax	Max LabCount trong năm
LabCountMin	Min LabCount trong năm
LabCountAvg	Avg LabCount trong năm
LabCountStd	Std LabCount trong năm
LabCountSum	Sum LabCount trong năm
LabClaimCount	Số lượng claim của LabCount trong năm
DrugCountMax	Max DrugCount trong năm
DrugCountMin	Min DrugCount trong năm
DrugCountAvg	Avg DrugCount trong năm
DrugCountStd	Std DrugCount trong năm
DrugCountSum	Sum DrugCount trong năm
DrugClaimCount	Số lượng claim của DrugCount trong năm

1.3.2 Strategy B

Trong chiến lược này, ta sẽ kết hợp các thuộc tính ở từng năm thành một bộ dữ liệu duy nhất. Sau đó sẽ chia tập training và testing theo tỉ lệ 8:2.

Các feature trong Strategy B:

Feature	Description
MemberID	Chỉ dùng làm index
AgeAtFirstClaim	Tuổi lúc claim lần đầu
Sex	Giới tính
ClaimCount	Số lượng claim toàn lịch sử
ClaimCountLatestYear	Số lượng claim Y2
ProviderCount	Số lượng Provider duy nhất toàn lịch sử
ProviderCountLatestYear	Số lượng Provider duy nhất trong Y2
VendorCount	Số lượng Vendor duy nhất toàn lịch sử
VendorCountLatestYear	Số lượng Vendor duy nhất trong Y2
PCPCount	Số lượng PCP duy nhất toàn lịch sử
PCPCountLatestYear	Số lượng PCP duy nhất trong Y2
SpecialtyCount_<value>	Số lượng claim từng giá trị toàn lịch sử
PlaceSvcCount_<value>	Số lượng claim từng giá trị toàn lịch sử
PayDelayMin	Min PayDelay toàn lịch sử
PayDelayMax	Max PayDelay toàn lịch sử
PayDelayAvg	Avg PayDelay toàn lịch sử
PayDelayStd	Std PayDelay toàn lịch sử
PayDelaySum	Sum PayDelay toàn lịch sử
PayDelayMinLatestYear	Min PayDelay trong Y2
PayDelayMaxLatestYear	Max PayDelay trong Y2
PayDelayAvgLatestYear	Avg PayDelay trong Y2

PayDelayStdLatestYear	Std PayDelay trong Y2
PayDelaySumLatestYear	Sum PayDelay trong Y2
LengthOfStayMin	Min LengthOfStay toàn lịch sử
LengthOfStayMax	Max LengthOfStay toàn lịch sử
LengthOfStayAvg	Avg LengthOfStay toàn lịch sử
LengthOfStayStd	Std LengthOfStay toàn lịch sử
LengthOfStaySum	Sum LengthOfStay toàn lịch sử
LengthOfStayCountNan	Số lượng LOS là NaN toàn lịch sử
LengthOfStaySumLatestYear	Sum LOS trong Y2
LengthOfStayMaxLatestYear	Max LOS trong Y2
LengthOfStayCountNanLatestYear	Số lượng LengthOfStay là NaN trong Y2
DSFSMin	Min DSFS toàn lịch sử
DSFSMax	Max DSFS toàn lịch sử
DSFSMaxLatestYear	Max DSFS trong Y2
CharlsonIndexMin	Min CharlsonIndex toàn lịch sử
CharlsonIndexMax	Max CharlsonIndex toàn lịch sử
CharlsonIndexAvg	Avg CharlsonIndex toàn lịch sử
CharlsonIndexMaxLatestYear	Max CharlsonIndex trong Y2
PrimaryConditionGroupCount_<value>	Số lượng claim từng giá trị toàn lịch sử
ProcedureGroupCount_<value>	Số lượng claim từng giá trị toàn lịch sử
LabCountSum	Sum LabCount toàn lịch sử
LabClaimCount	Số lượng claim của LabCount toàn lịch sử
LabCountSumLatestYear	Sum LabCount trong Y2
LabClaimCountLatestYear	Số lượng claim của LabCount trong Y2
DrugCountSum	Sum DrugCount toàn lịch sử
DrugClaimCount	Số lượng claim của DrugCount toàn lịch sử
DrugCountSumLatestYear	Sum DrugCount trong Y2
DrugClaimCountLatestYear	Số lượng claim của DrugCount trong Y2

Chapter 2

Predictive Modelling and Evaluation

2.1 Metrics

Metric được lựa chọn để đánh giá các model là:

$$RMSLE(predicted, actual) = \sqrt{\frac{1}{n}\sum_{i}^{n}[\log(predicted_i + 1) - \log(actual_i + 1)]^2}$$

Trong đó:
i là dòng dữ liệu thứ i trong tập testing
n là tổng số dòng dữ liệu trong tập testing
$predicted_i$ là kết quả dự đoán của dòng thứ i
$actual_i$ là kết quả thực tế của dòng thứ i

2.2 Algorithms

Trong phạm vi project này sử dụng 2 thuật toán XGBoost và Support Vector Regression để xây dựng mô hình hồi quy cho cả 2 chiến lược Strategy A và Strategy B, từ đó tìm ra được thuật toán và chiến lược nào là phù hợp nhất.

Các hyperparameter tốt nhất được tìm ra bằng cách sử dụng Grid Search với 5-fold Cross-validation.

Do hạn chế về mặt thời gian làm project cũng như thời gian training tất cả dữ liệu quá lâu, project này chỉ sử dụng 5.000 dòng dữ liệu để làm hyperparameter tuning.

Sau đó dùng tham số tốt nhất tìm được để train lại với tất cả dữ liệu và test lại với tập test để có số liệu cuối cùng.

2.2.1 XGBoost

Các hyperparameter sẽ được tuning là:

- **learning_rate**: Khảo sát các giá trị: 0.01, 0.1, 1.
- **max_depth**: Khảo sát các giá trị: 3, 5, 7.
- **subsample**: Khảo sát các giá trị: 0.3, 0.5, 0.7.
- **n_estimators**: Khảo sát các giá trị: 100, 300, 500

Kết quả tìm được như sau:

	Strategy A	Strategy B
Params	abc	abc
Best score	vv	vv
Testing set score	cc	cc

2.2.2 Support Vector Regression

Do giới hạn về thời gian và kinh nghiệm, trong project này chỉ sử dụng kernel **rbf**. Các hyperparameter sẽ được tuning là:

- **C**: Khảo sát các giá trị: 0.01, 0.1, 1, 10
- **epsilon**: Khảo sát các giá trị: 0.0001, 0.001, 0.01, 0.1, 1
- **gamma**: Khảo sát các giá trị: 'auto', 'scale'

Kết quả tìm được như bên dưới:

	Strategy A	Strategy B
Params	abc	abc
Best score	vv	vv
Testing set score	cc	cc

Chapter 3

Conclusion

3.1 Đánh giá hạn chế

- Hiện tại chỉ dùng 10.000 dòng dữ liệu để huấn luyện mô hình do thời gian train quá lâu nên kết quả mô hình chưa phải kết quả cuối cùng.
- Chưa có phương pháp để loại bỏ nhiễu và outlier.
- Chưa đủ domain knowledge để đánh giá mức độ quan trọng của các feature, nhất là các feature: Specialty, PrimaryConditionGroup và ProcedureGroup.

3.2 Future work

- Cố gắng giảm số chiều dữ liệu hoặc loại bỏ outlier để train nhanh hơn.
- Thử nghiệm tinh chỉnh các tham số khác của các thuật toán để tìm ra mô hình có độ chính xác cao hơn.
- Thử nghiệm các phương pháp khác như Ensemble Learning.