머신러닝 과제 보고서

- 4. 응급실 방문 예측 모델 -

제작자 : 김정호
연락처 : 010-7381-0416
Github address : https://github.com/kjh123qw/homework
e-mail : kjh123qwa@gmail.com

목차

1. 사용할 데이터 및 Target Label 추출

- 전처리 후 학습에 사용한 데이터 통계

데이터 수	41,810	Train Data	37,629	90%	Label:0*	35,565	85.0%
					Label:1*	2,064	5.0%
		Test Data	4,181	10%	Label:0*	3,952	9.5%
					Label:1*	229	0.5%

Label:0 => 30일 이내 응급실 방문하지 않은 환자 / Label:1 => 30일 이내 응급실 방문한 환자

2. 검증

- 선택한 검증 방법
  - StaratifiedKFold를 사용한 Cross Validation 학습
- 선택한 Metric
  - Logloss
  - Precision(정밀도)
  - Recall(재현율)
  - F1 Score
  - AUC(Area Under the ROC Curve)

3. 모델 및 전처리(1)

- 선택한 모델
  - XGBoost Classifier
- 전처리 방법
  - enc_visit_concept_id : 병원 방문 유형을 0,1,2로 인코딩한 데이터입니다.
    - 기존 visit_concept_id 데이터를 순서대로 0,1,2로 인코딩 한 데이터입니다.
  - enc_gender_concept_id : 환자의 성별 유형을 0,1로 인코딩한 데이터입니다.
  - visit_days : 병원 방문 기간(days, 소수점 포함)입니다.
    - 방문 데이터의 방문 종료 날짜와 시작 날짜를 datetime 객체로 변환한 뒤 방문 종료 날짜에서 시작 날짜를 빼서 구했고, 시간은 소수로 표현하였습니다. ex) 0.01, 10.4
  - visit_age : 병원 방문 당시 나이입니다.
    - 환자 데이터에서 birth_datetime와 병원 방문 종료 시점의 날짜의 차이를 통해 구한 값입니다. 일 수로 구한 값에 365를 나누어 구했습니다.

3. 모델 및 전처리(2)

- 전처리 방법
  - condition_concept_count : 병원 방문 종료 시점에 갖고있는 질병의 수를 표현한 데이터입니다.
    - 병원 방문 종료 시점을 기준으로 진단(병명) 데이터를 확인하여 환자가 당시 갖고있는 병의 수를 나타냈습니다.
  - top6_disease_check : 30일 이내에 응급실을 방문한 환자들이 당시 앓고있는 있는 진단(병명)을 추출하여 각 질병마다 count 한 뒤 가장 많은 순으로 정렬한 6개의 질병을 Target 환자가 병원 방문 종료 시점에 진단 받아 있는지 여부를 나타냅니다.
    - 30일 이내에 응급실을 방문한 환자들로부터 가장 많이 앓고있던 6개의 진단(병명)을 구한 뒤 환자 방문 종료 시점을 기준으로 앓고 있는 병명 중 해당 6개의 병명 코드가 존재하는지 여부를 확인하여 Y:1/N:0으로 나타냈습니다.

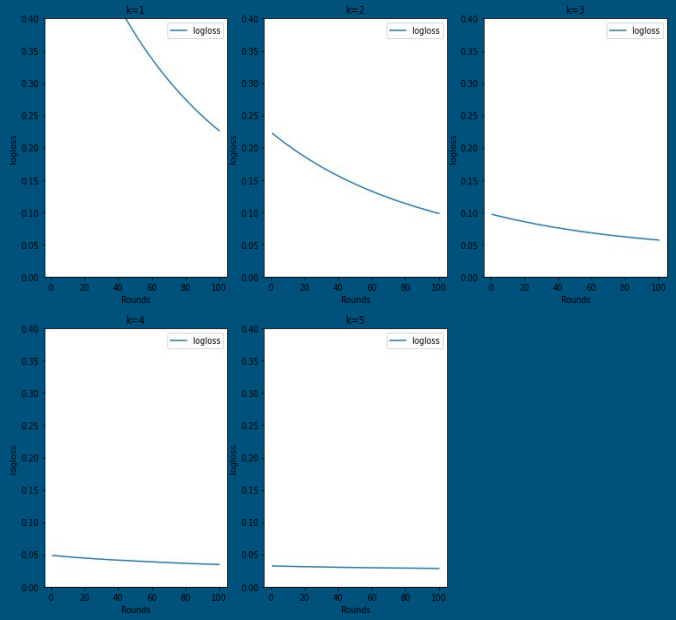
3. 모델 및 전처리(3)

- 전처리 방법
  - drug_concept_count : 병원 방문 종료 시점에 처방 돼 있는 의약품의 수를 표현한 데이터입니다.
    - 병원 방문 종료 시점을 기준으로 의약품 처방 데이터를 확인하여 환자가 당시 처방받은 의약품의 수를 나타냈습니다.
  - top9_drug_check : 30일 이내에 응급실을 방문한 환자들이 당시 처방 되어 있던 의약품을 추출하여 각 의약품마다 count 한 뒤 가장 많은 순으로 정렬한 9개의 의약품을 Target 환자가 병원 방문 종료 시점에 처방 받아 있는지 여부를 나타냅니다.
    - 30일 이내에 응급실을 방문한 환자들로부터 가장 많이 처방받고 있었던 9개의 의약품을 구한 뒤 Target 환자의 방문 종료 시점을 기준으로 처방 받고 있는 의약품 중 해당 9개의 의약품 코드가 존재하는지 여부를 확인하여 Y:1/N:0으로 나타냈습니다.

4. 실험 내용 – 하이퍼 패러미터 변경 실험(1)

옵션 : 기본	
StratifiedKFold의 n_split	5
EarlyStopping의 rounds	3
XGBClassifier의 metric_name	logloss
XGBClassifier의 n_estimators	100
XGBClassifier의 max_depth	5
XGBClassifier의 learning_rate	0.01

- Logloss 그래프



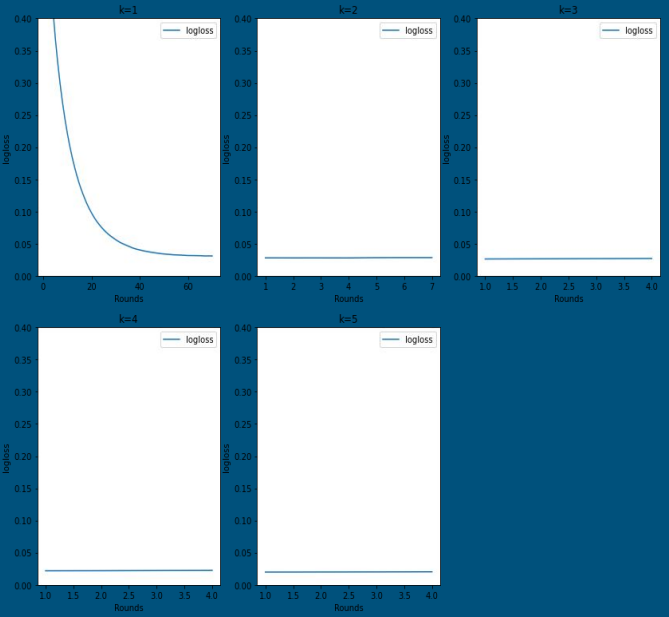
- Test data 결과

----------Test----------	----------Test----------
테스트 데이터 수	4181
테스트 데이터 정확도 / 수	94.5% / 3949
Logloss	0.148
Precision	0.485
Recall	0.210
F1 Score	0.293
AUC	0.598
테스트 수 [label:0]	3952
테스트 정확도 / 수 [label:0]	98.7% / 3901
테스트 수 [label:1]	229
테스트 정확도 / 수 [label:1]	21.0% / 48

4. 실험 내용 – 하이퍼 패러미터 변경 실험(2)

옵션 : 1	
StratifiedKFold의 n_split	5
EarlyStopping의 rounds	3
XGBClassifier의 metric_name	logloss
XGBClassifier의 n_estimators	100
XGBClassifier의 max_depth	5
XGBClassifier의 learning_rate	0.1

- Logloss 그래프



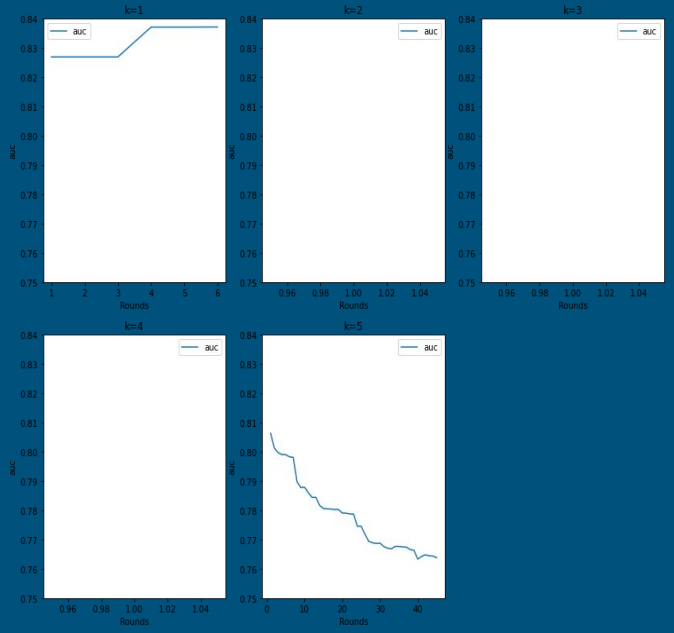
- Test data 결과

----------Test----------	----------Test----------
테스트 데이터 수	4181
테스트 데이터 정확도 / 수	94.5% / 3953
Logloss	0.138
Precision	0.505
Recall	0.236
F1 Score	0.321
AUC	0.611
테스트 수 [label:0]	3952
테스트 정확도 / 수 [label:0]	98.7% / 3899
테스트 수 [label:1]	229
테스트 정확도 / 수 [label:1]	23.6% / 54

4. 실험 내용 – 하이퍼 패러미터 변경 실험(3)

옵션 : 2	
StratifiedKFold의 n_split	5
EarlyStopping의 rounds	5
XGBClassifier의 metric_name	auc
XGBClassifier의 n_estimators	100
XGBClassifier의 max_depth	5
XGBClassifier의 learning_rate	0.01

- AUC 그래프



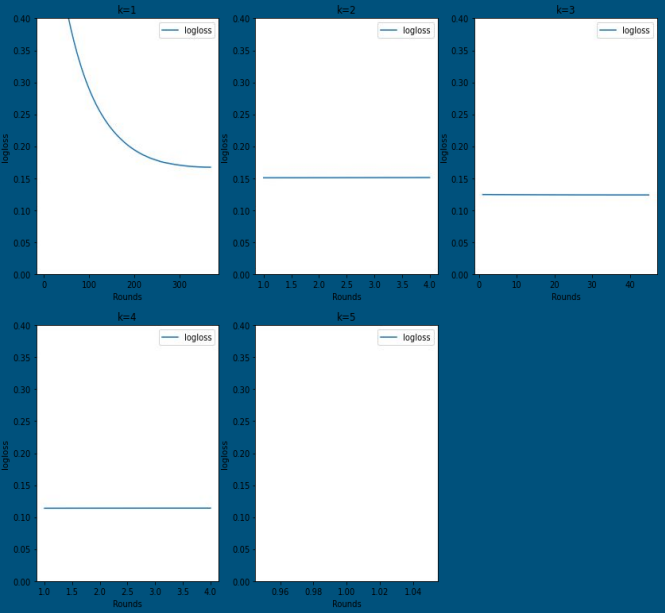
- Test data 결과

----------Test----------	----------Test----------
테스트 데이터 수	4181
테스트 데이터 정확도 / 수	94.4% / 3947
Logloss	0.142
Precision	0.475
Recall	0.205
F1 Score	0.287
AUC	0.596
테스트 수 [label:0]	3952
테스트 정확도 / 수 [label:0]	98.7% / 3900
테스트 수 [label:1]	229
테스트 정확도 / 수 [label:1]	20.5% / 47

4. 실험 내용 – 하이퍼 패러미터 변경 실험(4)

옵션 : 3	
StratifiedKFold의 n_split	5
EarlyStopping의 rounds	3
XGBClassifier의 metric_name	logloss
XGBClassifier의 n_estimators	500
XGBClassifier의 max_depth	5
XGBClassifier의 learning_rate	0.01

- Logloss 그래프



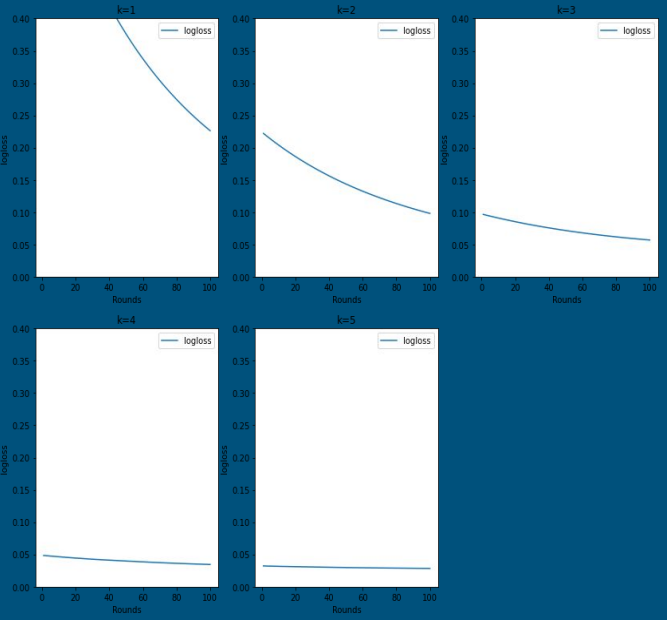
- Test data 결과

----------Test----------	----------Test----------
테스트 데이터 수	4181
테스트 데이터 정확도 / 수	94.5% / 3953
Logloss	0.143
Precision	0.505
Recall	0.236
F1 Score	0.321
AUC	0.611
테스트 수 [label:0]	3952
테스트 정확도 / 수 [label:0]	98.7% / 3899
테스트 수 [label:1]	229
테스트 정확도 / 수 [label:1]	23.6% / 54

4. 실험 내용 – 입력 데이터 변경 실험(1)

입력 데이터 조합 : 기본	
enc_visit_concept_id	Y
enc_gender_concept_id	Y
visit_days	Y
visit_age	Y
condition_concept_count	Y
top6_disease_check	Y
drug_concept_count	Y
top9_drug_check	Y

- Logloss 그래프



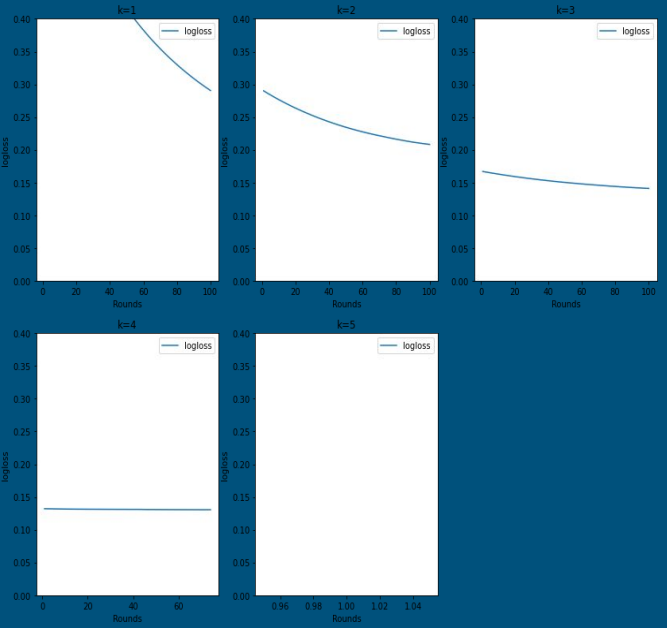
- Test data 결과

----------Test----------	----------Test----------
테스트 데이터 수	4181
테스트 데이터 정확도 / 수	94.5% / 3949
Logloss	0.148
Precision	0.485
Recall	0.210
F1 Score	0.293
AUC	0.598
테스트 수 [label:0]	3952
테스트 정확도 / 수 [label:0]	98.7% / 3901
테스트 수 [label:1]	229
테스트 정확도 / 수 [label:1]	21.0% / 48

4. 실험 내용 – 입력 데이터 변경 실험(1)

입력 데이터 조합 : 1	
enc_visit_concept_id	N
enc_gender_concept_id	N
visit_days	Y
visit_age	Y
condition_concept_count	Y
top6_disease_check	Y
drug_concept_count	Y
top9_drug_check	Y

- Logloss 그래프



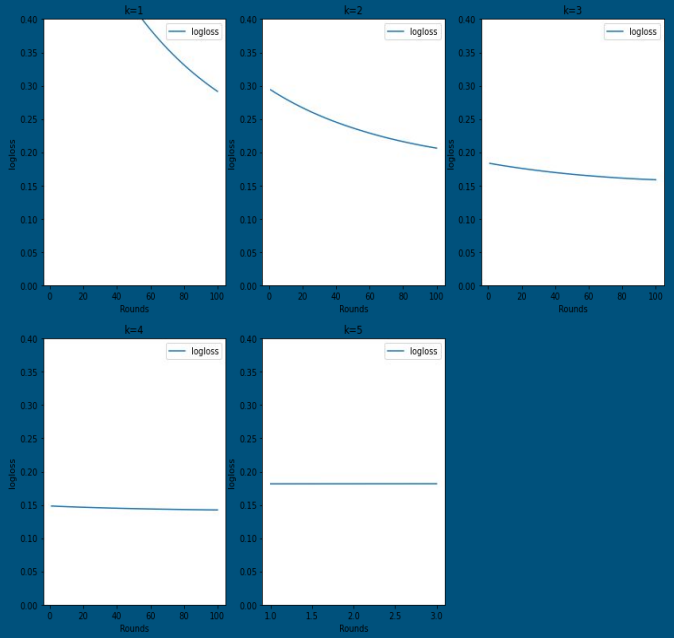
- Test data 결과

----------Test----------	----------Test----------
테스트 데이터 수	4181
테스트 데이터 정확도 / 수	94.8% / 3963
Logloss	0.158
Precision	0.789
Recall	0.066
F1 Score	0.121
AUC	0.532
테스트 수 [label:0]	3952
테스트 정확도 / 수 [label:0]	99.9% / 3948
테스트 수 [label:1]	229
테스트 정확도 / 수 [label:1]	6.6% / 15

4. 실험 내용 – 입력 데이터 변경 실험(2)

입력 데이터 조합 : 2	
enc_visit_concept_id	Y
enc_gender_concept_id	Y
visit_days	N
visit_age	N
condition_concept_count	Y
top6_disease_check	Y
drug_concept_count	Y
top9_drug_check	Y

- Logloss 그래프



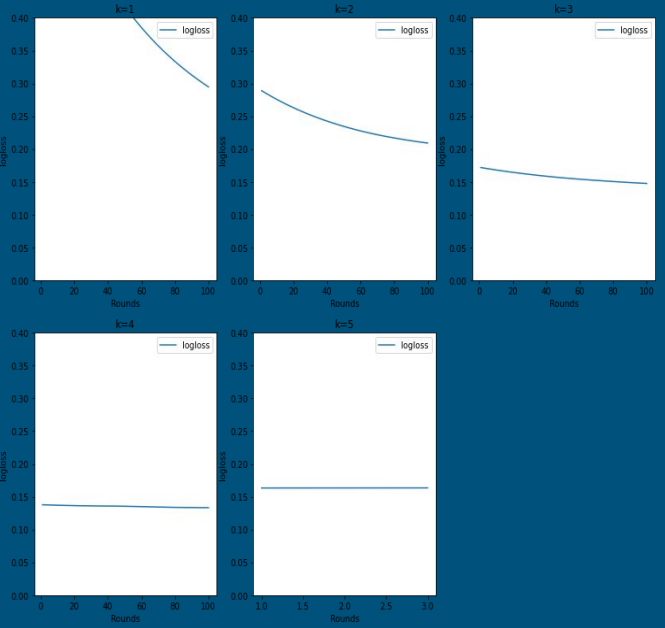
- Test data 결과

----------Test----------	----------Test----------
테스트 데이터 수	4181
테스트 데이터 정확도 / 수	94.5% / 3953
Logloss	0.152
Precision	0.556
Recall	0.022
F1 Score	0.042
AUC	0.510
테스트 수 [label:0]	3952
테스트 정확도 / 수 [label:0]	99.9% / 3948
테스트 수 [label:1]	229
테스트 정확도 / 수 [label:1]	2.2% / 5

4. 실험 내용 – 입력 데이터 변경 실험(3)

입력 데이터 조합 : 3	
enc_visit_concept_id	Y
enc_gender_concept_id	Y
visit_days	Y
visit_age	Y
condition_concept_count	N
top6_disease_check	N
drug_concept_count	Y
top9_drug_check	Y

- Logloss 그래프



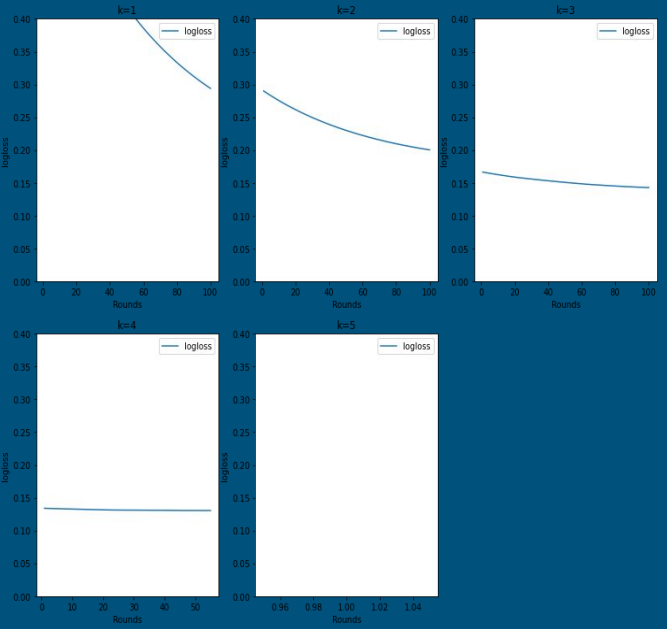
- Test data 결과

----------Test----------	----------Test----------
테스트 데이터 수	4181
테스트 데이터 정확도 / 수	94.4% / 3947
Logloss	0.157
Precision	0.473
Recall	0.188
F1 Score	0.269
AUC	0.588
테스트 수 [label:0]	3952
테스트 정확도 / 수 [label:0]	98.8% / 3904
테스트 수 [label:1]	229
테스트 정확도 / 수 [label:1]	18.8% / 43

4. 실험 내용 – 입력 데이터 변경 실험(4)

입력 데이터 조합 : 4	
enc_visit_concept_id	Y
enc_gender_concept_id	Y
visit_days	Y
visit_age	Y
condition_concept_count	Y
top6_disease_check	Y
drug_concept_count	N
top9_drug_check	N

- Logloss 그래프



- Test data 결과

----------Test----------	----------Test----------
테스트 데이터 수	4181
테스트 데이터 정확도 / 수	94.2% / 3939
Logloss	0.160
Precision	0.406
Recall	0.122
F1 Score	0.188
AUC	0.556
테스트 수 [label:0]	3952
테스트 정확도 / 수 [label:0]	99.0% / 3911
테스트 수 [label:1]	229
테스트 정확도 / 수 [label:1]	12.2% / 28

5. 후기

- 시간이 생각보다 더 부족하였습니다. 제출 1시간 전까지 label data를 잘못 된 곳에 대입하여 정확도가 90%가 나오고 있었는데, 이것을 생각하지 못하고 계속 실험을 하였습니다. 실험 중 방문 데이터를 빼고 학습을 시켰을 때 5%의 정확도가 나오는 것을 보고 정말 잘못됨을 감지하여 잘 생각해보니 label data를 잘못 넣었을 것 같다는 생각을 하게 되었습니다. 이후 전처리 완료하고 다시 실험을 하였습니다.
- 시간이 좀 더 있었다면, LGBM과의 성능 비교, 최근 30일 이내 병원 방문 횟수, 최근 1년간 병원 방문 횟수, 최근 3년간 입원 횟수 등의 파생 변수를 생성하여 예측해보고 싶습니다.