# 진료내역정보를 통한 수진자 병원비 예측

이름 강민진
학번 201524404
학과 정보컴퓨터공학부
학교 e-mail minj10092@pusan.ac.kr

## I. 문제설명(E)

건강보험공단에서 제공하는 진료내역정보(성별, 나이, 주소, 병명, 입원일 등) 여러 사항을 고려하여 수진자(서울시)의 병원비(본인 부담금)을 예측한다. 성능 향상을 통해 **보험사의 보험료 책정**에 지침이 될 수 있다. 또한, 시민들이 스스로 병원비가 얼마나 나올지를 책정해보는데 도움이 될 수 있다.

## II. 문제해결 방법

### A. *랜덤포레스트(Random Forest, Regression)*

하이퍼파라미터를 조정하지 않은 **기본 모델**로 랜덤 포레스트를 먼저 학습해서 베이스 라인으로 활용하였다. 교차 검증으로 5개로 쪼개서 성능을 평가한 후 0.64로 만족스러운 결과가 나오지 않았다 .

성능을 높이기 위해 하이퍼파라미터를 임의로 몇 가지 추려 **그리드 탐색(Grid Search)**을 적용하고 가장 좋은 하이퍼파라미터를 적용해보았다. 하지만 교차 검증 평균 성능 0.7정도가 나오는 것을 확인하여 하이퍼파라미터 후보군을 몇 가지 더 추가하여 그리드 탐색을 학습시켜보았다. 하지만, 모든 경우의 수를 점검하는 그리드 탐색의 특성 상 시간이 매우 오래 걸려 원하는 시간 내에 결과를 얻을 수 없겠다는 생각이 들었다.

적절한 하이퍼파라미터를 찾는 탐색 방법들을 찾다보니 랜덤하게 뽑아서 시간을 본인이 iteration을 설정하여 조정할 수 있는 **랜덤 그리드 탐색(Randomized Grid Search)**을 발견하고 적용해보았다. 앞서 수행한 그리드 탐색으로 발견한 괜찮은 하이퍼파라미터들을 후보군으로 넣어 활용하였더니 평균 성능 0.71로 약간의 성능을 향상시킬 수 있었다.

### B. *SVM(Support Vector Machine, Regression)*

하이퍼파라미터를 조정하지 않은 모델(모두 default)로 SVM을 먼저 학습해서 베이스 라인으로 활용하려고 했으나, 데이터가 많았는지 학습이 몇시간이 지나도록 끝이 나지 않았다. 그래서 kernel = 'rbf', gamma = 'scale', max_iter = 1000, C = 0.3으로 임의로 만들어준 것을 베이스 라인으로 활용하였다. 평균 성능은 -0.05가 나왔다. 학습이 제대로 되지 않은 것을 확인할 수 있었다.

성능 향상을 위해 Bayes Search 을 적용하여 학습을 시켰으나 복잡하게 조건을 넣자 학습 시간이 매우 많이 소요되었으며 단순하게 넣자 약 -0.01 로 원하는 정도로 성능이 향상되지 않았다.

## III. 실험 내용

### A. *실험 절차*

1) 공공데이터포털에서 진료내역정보(출처 : 건강보험공단) 데이터를 수집

2) 데이터의 항목별 세부사항을 확인 후 어떤 정보를 활용할지 결정(성별, 연령대, 시도(서울시), 서식 코드, 진료 과목 코드, 주상병(주병) 코드, 요양일수, 입내원일수, 심결 가산율(병원 수준에 따른 진료비 가산 비율), **병원비**(심결 본인 부담금), 총 처방일수)

3) 사용하지 않을 Feature 와 결측치를 제거

4) 사용할 Feature 타입 변환 (int, float, categorical)

5) 전체 데이터를 활용하기에는 RAM 용량이 부족하다고 판단하여 서울시만 특정하였고, 각 질병에 해당하는 것이 3000 개 이상인 것들만 활용(one-hot encoding 시 데이터가 너무 늘어나 학습이 어렵다고 판단하였으며, 적은 표본이 현실을 대표할 수 없다고 판단하여 제거)

6) Categorical feature 에 one-hot encoding 적용

7) 병원비(EDEC_SBRDN_AMT)와 상관관계를 계산(양의 상관관계 및 음의 상관관계)

8) 머신러닝을 위한 데이터 전처리

- train/test set 으로 분할
- 데이터 스케일링(Standard scaler 적용)

9) 학습 및 튜닝(Random Forest Regressor)

- 기본 모델(Baseline)
- 그리드 탐색(Grid Search)
- 랜덤 그리드 탐색(Randomized Grid Search)

10) Feature Importance 확인

11) 학습 및 튜닝(SVM)

- 기본 모델(Baseline)
- Bayes Search

### B. *측정 결과(cv = 5, 5 개의 평균 성능, random_state = 42 로 설정해 동일한 Train/Test 로 학습하였다.)*

**랜덤 포레스트**

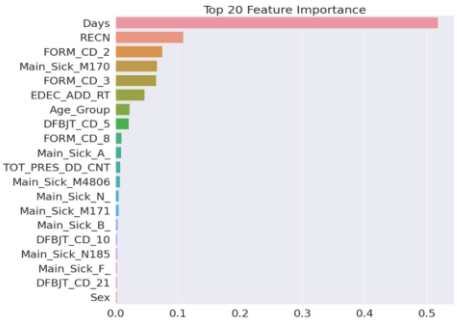
- 모델(베이스라인) : 약 0.64
- 그리드 탐색 적용 : 약 0.7
- 랜덤 그리드 탐색 : 약 0.71

SVM

- 기본 모델(베이스라인) : 약 -0.05
- Bayes 탐색 적용 : 약 -0.03(단순하게 돌린 결과이다. 많은 후보군을 넣자 5 시간을 돌려도 학습 결과가 나오지 않았다.)

*C. 측정 해석*

랜덤 포레스트 학습 결과 n_estimators = 40, max_depth = 12, min_samples_leaf = 2, min_samples_split = 20, max_features = 40 으로 학습한 모델이 가장 좋은 성능(0.71)을 보이는 것을 확인할 수 있었다.




[그림 1] Top 20 Feature Importance

Feature_importance 를 확인한 결과 Days(입원 및 병원 내 투약)이 **0.51** 으로 분기에 핵심 요소임을 파악할 수 있었다. 그 뒤로

- RECN(입원일수, 0.1)
- FORM_CD_2(의과 입원, 0.07)
- M170(병명 : 양쪽 일차성 무릎 관절증, 0.07)
- FORM_CD_3(의과 외래, 0.06)
- EDEC_ADD_RT(심결 가산율(병원 수준), 0.05)
- Age_Group(나이, 0.02)

순으로 분기를 결정하는 요인임을 확인할 수 있다.

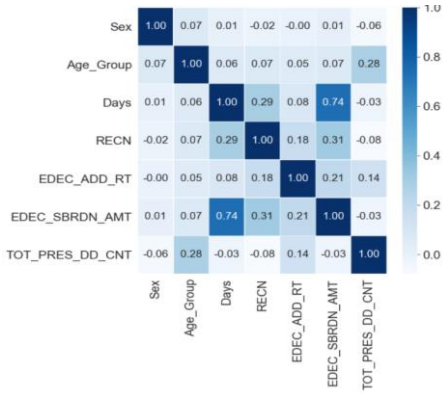
## IV. 결과

의미를 도출하기 위해 연관관계(상관관계)를 활용하였다. 주요 6 가지 특징부터 고려하였다.

- **입원일 수와 병원내 투약일 수**(Days, 0.74), 입원일 수(RECN, 0.31)가 가장 큰 영향을 미치는 것을 확인할 수 있다.

- 병원 등급(EDEC_SBRDN_AMT, 0.21), 나이(Sex, 0.07)는 적지만 병원비에 영향이 있는 것을 확인할 수 있다.

- 성별(Sex), 처방전에 따라 투약하도록 한 투약일 수(TOT_PRES_DD_CNT)는 병원비와 크게 연관성이 없음을 파악할 수 있다.




[그림 2] 병원비(본인 부담금, EDEC_SBRDN_AMT)와 주요 6 가지 특징과의 상관관계

또한 진료과목이 정신과, 신경과(정신적인 문제는 대부분 **입원**과 직결되기 때문이라고 유추) 순으로 양의 상관관계가 높게 나오는 것을 확인할 수 있었으며, 의과 **외래**(24 시간 미만 입원하는 환자)와 관절염, 급성 기관지염([4] 참고)는 입원을 하지 않으므로 병원비와 음의 상관관계가 높게 나오는 것을 확인할 수 있었다.

**성별, 처방일 수는 거의 병원비와 연관성이 없으며, 병원 수준, 나이에서 병원비가 증가되는 것보다는 입원과 직결되는 병, 진료과목 등 입원의 영향이 더 크다는 것을 파악할 수 있다.**

## V. 회고

데이터를 많이 줄여 학습에 어려움이 없을 것이라고 생각했지만 노트북으로 학습을 수행하는 시간이 꽤나 오래 걸렸다. 특히 그리드 탐색의 경우 후보군을 그리 많이 넣지 않았는데도 모든 경우의 수를 고려하는데 시간이 오래 걸렸다.

분명 random_state 를 잘 고정했는데 최종본을 다시 처음부터 돌려보니 매번 다른 결과가 도출되었다. 평가를 측정하는 방법으로 사용했던 교차 검증에도 다른 방식으로 random_state 를 적용해야 하는 것을 늦게 파악했다. 다시 그리드 탐색 및 랜덤 그리드 탐색을 적용하기에는 시간이 촉박하여 모델의 결과와 보고서의 결과가 다르게 나올 수 있다.(고정하고 이후에 다시 돌려봐야 겠다.)

얼떨결에 그리드 탐색을 먼저 하고 랜덤 그리드 탐색을 수행했는데 생각보다 괜찮은 방법이었던 것 같다. 다음부터는 범위를 넓게 해서 랜덤 그리드 탐색을 수행하고 광범위한 것들 중에 괜찮은 것들을 뽑아낸 후에 구체적인 성능 향상을 위해서 그리드 탐색을 적용해보는 것도 좋은 방법일 것 같다는 생각이 들었다. 그리드 탐색과 랜덤 그리드 탐색을 효과적으로 활용하지 못했는데 다음에는 동일한 시간이 주어져도 더 나은 성능을 보이는 모델을 만들 수 있겠다는 확신을 가지게 되었다.

상관관계를 잘 시각화했다면 결론의 좋은 근거가 될 수 있었을텐데 시간과 실력이 부족하여 아쉬웠다.

## 참고문헌, 자료


[1] https://data.go.kr/data/15007115/fileData.do: 공공 데이터 포털 진료내역정보(2018) 데이터

[2] https://injo.tistory.com/30 : Feature importance 시각화 참고

[3] https://nhiss.nhis.or.kr/bd/ad/bdada038cv.do : 질병 코드

[4] https://m.health.chosun.com/svc/news_view.html?contid=2016022403365 : 기사(급성 기관지염은 대부분 외래 환자)