기상청

기상에 따른 혈관 질환 발생 예측 모델 개발

참 가 번 호	220119	팀 명	플라잉콘

Ⅰ. 서론

[심뇌혈관질환이란]

심뇌혈관질환이란 심혈관질환과 뇌혈관질환을 아우르는 용어로 심근경색증, 뇌졸중 등이 포함된다. 2020년 기준 통계청에 따르면 심장질환은 국내 사망 원인 2위이며, 뇌혈관 질환은 4위로 그 뒤를 이었다.

[날씨와 심뇌혈관질환의 관계]

이전부터 심뇌혈관질환의 발병이 기온, 대기오염과 같은 기상인자들과 연관이 있다는 연구결과들이 발표되어왔다(배현주, 2014, p346). 특히 심근경색증은 기온 및 습도와 관련성이 높은 것으로 보고되어 있으며 최근에는 기압과 평균기온의 일 변동성이 뇌졸중 발생에 유의한 영향을 나타낸다는 결과가 새로 보고되었다(안혜연, 2016, p.840). 또한 2003년부터 2010년까지 서울을 중심으로 기상청의 기상관측자료와 국립환경과학원의 대기오염 측정자료, 공단의 심혈관질환 발생건수 등을 종합 분석한 결과, 대기 중 미세먼지가 1㎍/㎥증가하면 심혈관질환으로 입원할 확률은 1.26% 증가하는 것으로 나타났다(이승운, 2020, p276).

[분석의 의의 및 목적]

심뇌혈관질환은 골든타임이 무엇보다도 중요한데, 심근경색의 경우 2시간 이내, 뇌졸중의 경우는 3시간 이내에는 치료받아야 한다. 따라서 일자별 심뇌혈관질환 입원 건수를 미리 예측하여 의료인원 및 치료 시스템을 준비하는 것이 중요하다. 심뇌혈관질환과 연관성이 높은 날씨 및 인구학적, 지리적 파생변수들을 이용해 백병원의 심뇌혈관질환자 입원 건수를 예측하는 것이 분석의 목적이다.

Ⅱ. 전처리 및 EDA

[전처리]

1) 데이터 생성

(1) 관측 데이터

관측데이터는 크게 AWS와 ASOS로 나뉜다. 본 분석에서는 그 중 ASOS를 사용했다. AWS는 비교적 산악지역이나 섬처럼 사람이 관측하기 어려운 곳에 설치하여 국지적인 위험기상현상을 실시간으로 감시한다. 반면 ASOS는 주로 기상관서에 설치되고, 그 지역의 현재 기상 실시간 제공 및 기상 예보에 활용되고 있다. 따라서 분석에서 예측할 입원 건수는 사람들이 거의 거주하지 않는 험준한 산악지역이나 섬의 날씨보다는 사람들이 실제로 거주하는 지역의 날씨에 영향을 받는다고 생각하여 ASOS를 사용했다. 기상 변수로는 기온, 기압, 습도, 강수량, 풍속, 적설 등을 사용했다.

(2) 예보 데이터

제공된 예보데이터 중 3시간 기온, 6시간 강수량, 6시간 적설, 12시간 강수량, 12시간 적설, 습도, 일 최고기온, 일 최저기온, 풍속 데이터를 이용하여 24시간 뒤와 48시간 뒤의 예보데이터를 생성했다. 생성한 변수는 24시간 뒤 강수량, 24시간 뒤 신적설, 24시간 뒤 기온, 48시간 뒤 기온, 24시간 뒤 습도, 48시간 뒤 습도, 24시간 뒤 최고기온, 24시간 뒤 풍속, 48시간 뒤 풍속이며 해당 변수들은 다음과 같은 방식으로 생성했다.

※ 24시간 뒤 예보 : 해당 일 23시 기준 +4, +10, +16, +22 예보 값의 평균 또는 합

※ 48시간 뒤 예보 : 해당 일 23시 기준 +28, +38, +40, +46 예보 값의 평균 또는 합

(3) 지수 데이터

기상청 날씨마루 HIVE에서 제공된 기상생활지수 데이터 중 체감온도(A03)와 자외선지수(A07), 보건생활지수 중 뇌졸중위험지수(D07)을 이용하여 당일, 24시간 뒤, 48시간 뒤 지수값의 평균값, 최소값, 최대값, 범위를 계산했다.

(4) 대기오염 데이터

대기오염 데이터는 에어 코리아의 시간별 데이터를 사용하였다. 시간별로 하루에 24번 기록된 데이터 값을 평균하여 일별 데이터로 사용하였다. 사용 변수는 미세먼지 (PM10), 오존(O3), 이산화질소(NO2), 일산화탄소(CO), 아황산가스(SO2) 수치이다.

2) 이상치 제거 및 결측치 처리

각 데이터의 변수는 EDA를 통해 확인한 후 이상치로 판단되는 값들을 모두 제거하였다. 예보 데이터, 관측 데이터, 지수 데이터, 대기 오염 데이터 모두 결측치가 존재했는데 이를 처리하는 방법으로는 MICE(Multivariate Imputation by Chained Equations)와 interpolate를 사용하였다. MICE는 결측치를 처리하는 데 각 복원 모델에 따라 대치를 진행하는 실용적인 방법이다. 관측 데이터, 지수 데이터, 대기오염 데이터는 MICE를 이용하여 결측치 처리를 했고, 상대적으로 변수가 적은 예보 데이터는 interpolate를 이용해 단일 변수로 결측치 처리를 진행했다.

세종은 2012년 7월에 출범한 신도시이기 때문에 많은 기상 데이터가 결 측이었다. 특히 세종의 ASOS는 2020년에서야 설치되었기 때문에 2012년부터 2016년까지 모든 기간의 관측 데이터가 모두 결 측이었다. 따라서 관측 데이터와 대기오염 데이터의 모든 관측 지점의 위치를 지도로 시각화한 뒤 세종 근방의 관측 지점들의 평균값으로 세종 지역의 결측치를 대체하였다.

3) 데이터 병합

시도별 혹은 읍면동(ex. 강릉 혹 갈말읍)으로 이루어져 있는 데이터들을 하나의 지역(ex. 강원)으로 합치는 과정에서 지역별 인구학적 특징을 날씨에 반영하기 위해 인구 가중치를 적용했다.

4) 변수 추가

(1) 기상 지연변수(lag1 ~ lag21)

기상변수별로 해당일 하루 전부터 21일 전까지 총 21개의 lag 변수를 생성했다.

(2) 범주형 변수 인코딩

여러 가지 방식의 인코딩 방식을 적용한 뒤 모델의 성능이 높은 방식을 사용했다.

방식 1) 원-핫 인코딩(One-Hor encoding)

지역, 성별, day1 등 범주형 변수의 경우 원-핫 인코딩을 사용하여 범주의 개수만큼 더미변수를 생성하였다.

방식 2) 주기 인코딩(Cyclical encoding)

요일과 계절은 계속 반복되는 주기를 가진 변수이므로, 주기성을 반영하여 주기 인코딩(cyclical encoding)을 진행했다. 요일과 계절은 각각 sin과 cos으로 극좌표계를 나타내는 2개의 변수들로 표현된다.

(3) 주요 기상변수(온도, 압력, 강수량, 풍속)의 7일 평균 및 14일 평균

(4) 일교차 변수 및 큰 일교차 횟수

큰 일교차 횟수 변수는 일주일간 일교차가 10도 이상인 날의 횟수를 값으로 넣었다.

(5) 심정지, 고혈압, 뇌졸중 약물치료 및 진단, 가족력

지역별·성별·연도별 심정지, 고혈압, 뇌졸중 약물치료 환자 수 합, 진단 환자 수 합, 가족력 여부 변수를 추가했다.

(6) 월별 지역별 50대 이상 인구수

(7) 지역별 백병원 개수

해당 지역의 백병원 개수를 세어 범주형 변수로 만들었다. 거주지로부터 백병원과의 거리가 먼 경우 접근성이 떨어지기 때문에 입원건수에도 영향을 미칠 것이라 판단하여 해당 변수를 추가하였다.

(8) 1일여부, 월요일여부, 주말, 요일, year, 공휴일여부

(9) 지역별 연도별 음주율, 흡연율, 비만율, 신체활동량, 건강인지율

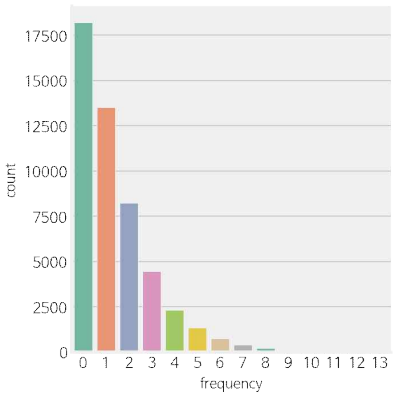
[EDA]

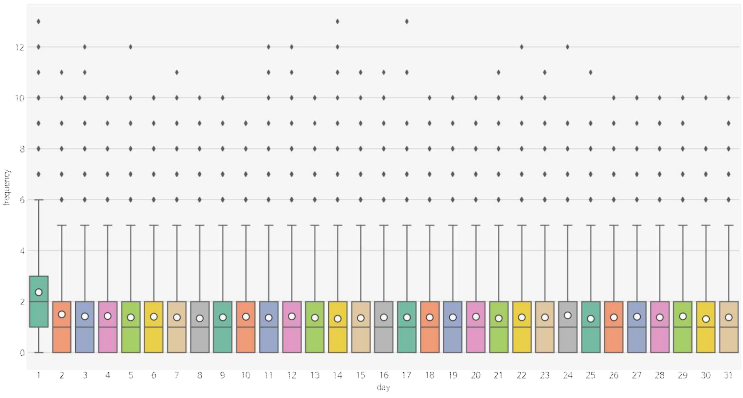
1) frequency의 분포

전체 frequency 값은 0부터 13까지 넓게 분포한다. 그 중 0이 압도적으로 많고, 다음으로 1과 2가 많았다. 5 이상의 값은 매우 드물었다.

2) 일별 frequency

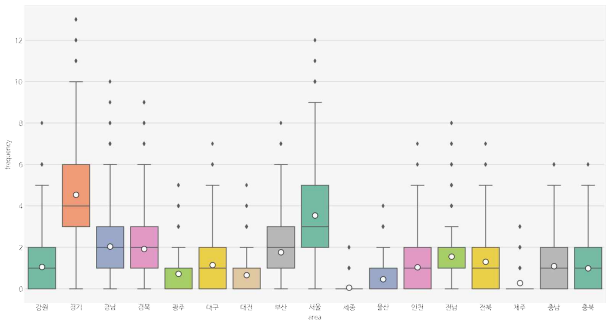
일별 frequency를 확인해보면, 1일이 유의하게 높음을 확인할 수 있다.

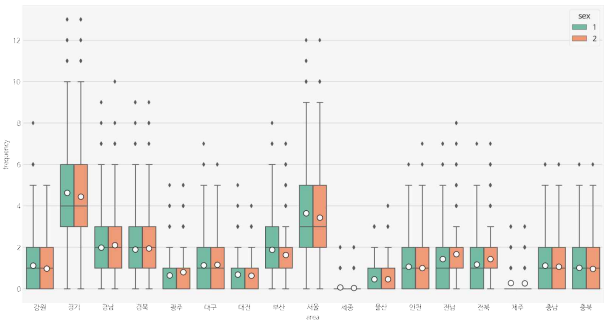




3) 지역별·성별 frequency

- frequency 값이 지역별로 차이가 크다. 대부분의 지역은 median이 1에서 2 사이의 값을 가지지만, 경기과 서울 지역은 3~4 정도로 매우 두드러지는 값을 가진다. 반면 세종과 제주 지역은 대부분의 data가 0값을 가지는 것을 볼 수 있다. 강원/대구/인천/충남/충북, 경남/경북, 광주/대전/울산, 전남/전북이 각각 비슷한 plot을 보인다.
- 대부분의 지역에서는 성별에 따른 차이가 거의 없었지만 부산, 전남, 전북과 같이 일부 다른 특성을 보이는 경우도 있었다.
- 따라서 EDA 결과, 우리는 모델에 지역별, 성별, 일자별 weight를 주기로 했다.





Ⅲ. 모델링

1) 개별 모형 설명

Modeling	shape	변수	설명
Bayesian Ridge	(49674, 247)	날씨변수 lag1-lag3 큰 일교차 횟수 지역별, 성별 음주 시도별 성별 신체활동량 지역별 고위험 음주율 지역별 비만율 지역별 주관적 건강인지율	지역을 factor로 넣은 모델링

Lasso Regression	(49674, 1686)	날씨변수 lag1~lag21 큰 일교차 횟수 지역별, 성별 음주 시도별 성별 신체활동 운동 지역별 고위험 음주율 지역별 비만율 지역별 주관적 건강인지율	지역별로 나누어 modeling
Random Forest	(49674, 229)	날씨변수 lag1-lag3	지역을 factor로 넣은 모델링
Bayesian Ridge 2주 학습	(49674, 261)	날씨변수 lag1-lag3 큰 일교차 횟수 지역별, 성별 음주 시도별 성별 신체활동 운동 지역별 고위험 음주율 지역별 비만율 지역별 주관적 건강인지율 4개 중요변수(50대이상 비율, 심장병_가족력, 뇌졸증_약물치료, 뇌졸중_진단)의 역수, 제곱, 로그 변수 frequency의 lag14	지역을 factor로 넣은 모델링, 2주 학습

2) 모형 앙상블

우선 Bayesian Ridge, Random Forest 모델을 5:5로 앙상블 하여 br_rf 모델을 생성한 후, br_rf, Lasso Regression, Bayesian Ridge_2_week 모델을 2:2:1로 앙상블 했다. 그리고 앞선 EDA 결과에서 알 수 있듯 frequency 값은 지역별, 성별, 일자별에 따라 서로 다른 분포를 보인다. 이 차이를 모델에 반영하기 위해 앙상블 한 모델에 지역별, 성별, 일자별 평균값을 가중치로 부여하여 최종 모델을 생성하였다.

3) 결과 해석

(1) 모형에 대한 해석

데이터가 충분하지 않은 경우, 머신러닝 및 딥러닝보다 structured modeling이 잘 작동하는 경우가 많다. 본 데이터는 지역별로 나누어진 데이터로 n의 개수가 충분하지 않아 모델 적용 결과 xgboost, lightgbm, catboost와 같은 부스팅 모형이나 lstm, dnn, tft 같은 시계열에 강한 딥러닝 모형보다 단순한 회귀 모형이 더 적절한 결과가 나와 가설과 일치했다.

(2) 중요 요인에 대한 해석

- 지역: 지역 및 인구학적 특징에 따라 frequency 값은 큰 편차를 보인다. 경기, 서울은 입원건수가 상대적으로 높았고 제주, 세종은 상대적으로 낮게 분포되어 있다.

- 일자: 일자별 frequency 값은 편차를 보이는데 특히 1일인 경우 다른 날들에 비해 frequency 값이 상대적으로 높았다. 이는 행정상 보험료 산출을 위해 익월 1일로 입원 일자가 처리되었기 때문이라고 추정할 수 있다.

Ⅳ. 활용방안 및 개선방향

[모델을 이용한 활용방안 및 기대효과]

기상 및 지리적, 인구학적 데이터를 기반으로 백병원의 심뇌혈관질환자 입원건수를 예측하는 모델을 병원 전산 시스템에 도입하여 일자별 환자 수 변화에 대응할 수 있다. 이는 높은 입원건수가 나타나는 계절, 기상요건 등에 따라 필요 인력 및 의료자원을 사전에 구축해 시의적절한 대응을 기대할 수 있다.

지역별 입원건수의 차이를 보이므로 모델 기반으로 지역 맞춤형 보건사업 및 정책을 수립하여 심뇌혈관질환 취약 지역에 지원을 강화해 지역사회 건강수준의 전반적인 향상을 기대할 수 있다.

[개선점]

본 팀은 기상데이터, 대기오염 데이터, 지역 및 인구학적 데이터를 활용하여 전국 백병원의 심뇌혈관질환자 입원 건수를 Bayesian Ridge, Random Forest, Lasso 등을 이용하여 예측하였다. 이는 결과적으로 좋은 예측 성능을 보였으나 분석 과정에서 판단한 개선할 점은 다음과 같다.

1) 분석 과정에서 지역(시도)이 중요한 변수로 밝혀졌으나 각 병원의 입원건수를 보다 정확하게 예측하기 위해서는 각 병원별 입원건수 데이터를 반영하는 것이 필요하다.

2) 골든타임을 놓치지 않는 것이 중요하기에 일자별 보다는 시간별 기상데이터 및 입원건수 데이터를 활용한다면 더 좋은 성능을 보일 것으로 기대된다. 또한, 지역(시도)별이 아닌 각 환자 개인의 음주여부, 흡연여부, 가족력 등의 데이터가 있다면 모델의 성능을 더욱 높일 수 있을 것이다.

3) 현재 기상데이터는 각 시도의 일부 읍면동 데이터만을 담고 있어서 시도 내 지역적 동질성을 가정하고 있다. 예를 들어, 강원도의 경우 내륙에 위치한 지역이 있는 반면, 해안에 위치한 지역 또한 있다. 이러한 산, 해안, 내륙 등의 지리적 특징을 담기 위해서는 보다 다양한 읍면동 데이터를 활용하는 것이 필요해 보인다.

V. 참고문헌

김영수, 박호정. (2021). DNN과 LSTM 활용한 일일 전력수요모델 개발 및 예측. 한국기후변화학회. 12(3), 241-253.

조동철, 한희일. (2021). 기상 데이터에서 대기 오염도 요소의 결측치 보완 기법 제안. 한국인터넷방송통신학회. 21(1). 181-187.

이승운, 정승권. (2020). 인천시 자치구별 미세먼지 농도에 따른 호흡기 및 심혈관계 외래환자수 상관분석. 한국환경보건학회지, 46(3), 276-284.

이승현, 여인권. (2020). 순환신경망을 이용한 질병발생건수 예측. 응용통계연구. 33(5), 627-637

안혜연, 정주희 외. (2016). 학술논문 분석을 통한 기상민감질환 선정 및 기상인자와의 관련성고찰. 한국환경학회지, 25(6), 839-851.

이선재, 여인권. (2016). 기상에 따른 고령환자의 질병 발생빈도 예측모형 비교. 응용통계연구. 29(1), 145-155

배현주. (2014). 서울시 미세먼지(PM10)와 초미세먼지(PM2.5)의 단기노출로 인한 사망영향. 한국환경보건학회지, 40(5), 346-354.

이세란, "한국 성인에서 심혈관질환 주요 위험 요인의 계절성에 대한 연구: 국민건강영양조사 제 4기 1차년도(2007) 자료를 이용하여". 석사학위논문, 연세대학교, 2009.