기상청

2022 날씨 빅데이터 콘테스트

참 가 번 호	220194	팀 명	주연과 조연들

I. 공모 배경 및 목표

1. 공모배경

2020년 기준 국내 사망 원인 2위는 심장 질환, 4위는 뇌혈관 질환으로 심뇌혈관질환은 국내 사망의 주요 원인이라 할 수 있다. 심뇌혈관질환의 진료비 및 관련 비용 또한 꾸준히 증가하는 추세이며, 해당 질환이 노년층에서 주로 발생하는 점을 감안하면, 인구고령화에 의한 사회경제적 부담도 늘어날 것으로 보인다.[i]

심뇌혈관질환의 주요 원인으로는 현대인의 생활 습관과 더불어 최근 기상 요인이 주목받고 있다. 관련 연구[ii]들 역시 선행되는 중으로, 심뇌혈관질환은 스스로 예방하기 어렵고 골든타임이 짧기 때문에 기상과 같은 주변 상황에 맞추어 조기검진을 통해 평소에 관리할 필요성이 있다.

2. 분석목표

본 공모안에서는 시도, 성별에 따른 심뇌혈관질환 발생 특성을 고려하여 머신러닝, 딥러닝 등 다양한 예측 모델을 구축한 후 최적의 모델을 찾는 것을 목표로 한다. 또한, XAI로 Black box 모델을 해석할 수 있게 하여 본 예측모델의 활용성을 높인다.

II. 활용데이터정의

1. 활용데이터

본 공모안에서는 날씨마루에서 제공한 날씨 데이터와 서울백병원·렉스소프트(주)에서 제공한 혈관질환 발생빈도 데이터를 사용하였다. 기상 요인에 따른 심뇌혈관질환 발생 관련 논문[iii]을 참고하여 제공 데이터 외에 다양한 요인들을 분석하기 위해 기상자료개방포털, 국립환경과학원으로부터 데이터를 추가로 수집했다.[1]

2. 활용변수

선행 연구[iv]에 기반하여 최종적으로 일별 평균/최저/최고기온, 평균상대습도, 평균/최저/최고 이슬점온도, 평균/최저/최고 해면기압, 상대습도예보를 활용변수로 선정했다.

III. 데이터 전처리

1. 결측치 처리

[1] '4. 활용데이터목록.xlsx' 참고

결측률이 50% 이상인 데이터는 분석에서 사용하지 않았다. 특정 관측소에 결측치가 발생한 경우 거리상 근처 5개 관측소 평균 혹은 선형보간법, 수정된 정규화비율 방법[v]을 사용하여 결측치를 대체하였다. 모든 관측소가 특정 기간에 대해 결측인 경우에는 선형보간법을 이용하거나 과거 관측치로 대체하였다.

2. 이상치 처리

기상청 '기상기후 품질검사 알고리즘'[vi]에 따라 이상치를 대체하였다. 또한 여름철 최저기온이 영하 30도인 것처럼 일반적인 범주를 벗어나는 경우 기상자료개방포털을 참조해 값을 대체하였으며, 미세먼지 데이터에 대해서는 1%로 윈저라이징을 실시하였다.

3. 일별 시도별 기상정보 대표값 산출

본 공모전의 예측 대상은 일별 시도별 성별 심뇌혈관질환 발생빈도(frequency)이므로 108개의 관측소가 가진 기상정보를 이용하여 17개 시도별 기상정보 대표값을 선정했다.

상세한 방법은 다음과 같다.

1. 각 관측소의 위도, 경도로부터 도로명 주소를 구한 후, 이를 통해 각 관측소가 속한 시도 추출
2. 일자와 시도가 같은 데이터에 대해 평균을 산출하여 이를 해당 일자의 시도 대표값으로 선정

2단계에서 사용할 수 있는 평균 산출 방법에는 단순평균, 관측소가 위치한 시군구의 인구수를 이용한 가중평균, 관측소별 거리를 고려한 가중평균 등이 있었다. 우선 관측소가 위치한 시군구의 개수가 적어 인구수를 이용한 가중평균을 구하는 것이 여의치 않았다. 여러 논문[vii]에서 기상정보 대표값 선정 시 단순평균을 사용했으며, 결과 해석이 용이하기 때문에 평균 산출 방법으로 단순평균을 선택하였다.

4. EDA

데이터 분석을 들어가기에 앞서 탐색적 데이터 분석(EDA)를 진행하였다. 요일, 월, 계절, 공휴일 등 시간의 흐름에 따라 심뇌혈관질환 발생빈도가 상이해질 것으로 보고, 각 시간 관련 변수에 따른 시도별 심뇌혈관질환 발생빈도의 분포를 확인하였다.

또한, 심뇌혈관질환 발생빈도의 자기상관성을 알아보기 위해 ACF를 확인해보니 경기지역은 7일 주기가 뚜렷하나, 다른 시도는 뚜렷한 주기성을 보이지 않았다.

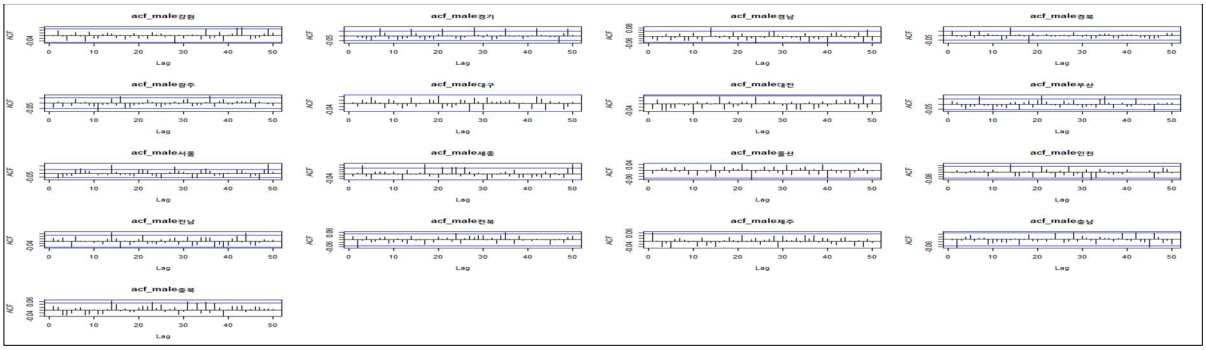

그림 1 시도별 ACF

5. 파생변수 생성

가. 시간 관련 변수

1. 요일: 월요일, 토요일, 일요일 더미변수에 대해 Chi-square test, fisher's exact test에서 유의한 차이를 보였으므로 각 요일에 대한 더미변수를 생성하였다.
2. 공휴일: 백병원이 일요일/공휴일 휴진임에 따라 공휴일 여부에 대한 더미변수를 생성하였다.
3. 월: 월초, 월말 등 월 중 며칠인지가 frequency에 영향을 미칠 수 있으므로 일자 변수를 만들었다.
4. 월, 계절 더미변수: 유의성 검정에서는 유의한 차이를 보이지 않았지만, 타 변수들과 joint하게 작용했을 때에는 frequency와 관계가 있기 때문에 포함시켰다.

나. 기상정보 관련 변수

1. 공기질: 각 기상 요소가 WHO 일일 기준치 초과여부를 나타내는 더미변수를 생성하였다. 생성 대상은 다음과 같다.
   (24시간기준) SO2, PM10, (8시간기준[2]) CO[3], O3, (1시간기준) NO2[4]
2. 폭염여부: 일 최고기온이 33도 이상이면 1, 그렇지 않으면 0을 부여했다.
3. 한파여부: 아침 최저기온(03:01~09:00)이 영하 12도 이하이면 1, 그렇지 않으면 0을 부여했다.
4. 기상정보별 N일 전 값: 각 기상정보에 대해 lag를 취하여 1일전~7일전 기상정보 변수를 파생시켰다. 이때 모든 변수에 대해 lag를 취할 경우 다중공선성의 문제가 있기 때문에 온도, 습도, 기압 변수 중에서 가장 변수 중요도가 높았던 일최저기온, 최저습도, 최저기압 세 변수들에 대해서만 lag 변수를 추가해주었다.

IV. 분석 기법 및 결과

시도별로 관측되는 기상정보와 인구 비율 등의 다른 특성을 가지고 있다는 점과 성별에 따른 심뇌혈관질환 발생빈도의 범위가 상이하다는 점을 고려하여 시도, 성별에 따라 다르게 예측모델을 구축했다.

전통적인 통계모형과 딥러닝, 머신러닝 예측모델을 생성한 뒤 교차검증(5-fold CV)을 수행하여 rmse가 가장 낮은 모델을 해당 시도, 성별에 대한 최적 모델로 선정하였다.

1. Baseline: 선형회귀분석

예측모델을 구축하기에 앞서 선형회귀분석을 이용하여 기준모델을 생성했다. '12~'15 데이터로 교차검증을 진행한 결과는 다음과 같다.

시도	강원	경기	경남	경북	광주	대구	대전	부산	서울	세종	울산	인천	전남	전북	제주	충남	충북
남성	1.269	2.279	1.519	1.467	0.866	1.193	0.805	1.387	2.027	0.273	0.717	1.054	1.227	1.050	0.566	1.062	1.117
여성	1.015	2.453	1.604	1.476	0.954	1.165	0.775	1.374	1.984	0.270	0.683	1.062	1.382	1.302	0.580	1.063	1.101

표 1 선형회귀분석 5-fold CV rmse

2 주 활동시간인 10시-18시 기준

3 WHO 기준이 존재하지 않으므로, 국내 기준 적용

4 15시 기준

2. 머신러닝 예측모델: LightGBM, XGB, RandomForest

가. RandomForest

RandomForest(이하 RF)는 여러 개의 decision tree로부터 최종 결과를 도출하는 모델로, 학습과정에서 자체적으로 변수선택을 진행한다는 장점이 있다.

본 공모안에서는 모델의 성능을 높이기 위해 variable importance 확인 및 관련 논문[viii]을 근거로 변수 선택 단계를 추가하고, 차원축소(PCA)도 진행하였으나 성능이 개선되는 모습을 보이지 않았다. 따라서 전체 변수를 사용하여 RandomForest 모델을 구축하였다.

하이퍼파라미터 튜닝은 최신 AutoML기법인 Optuna를 사용했다. 하이퍼파라미터를 튜닝하기 전과 후의 rmse를 비교해보았을 때, 평균 -2.4% 정도 감소함으로써 성능이 개선된 것을 확인할 수 있었다.

시도	강원	경기	경남	경북	광주	대구	대전	부산	서울	세종	울산	인천	전남	전북	제주	충남	충북
전	1.080	2.189	1.467	1.372	0.827	1.101	0.800	1.372	1.956	0.231	0.687	1.062	1.232	1.081	0.550	1.102	1.017
후	1.104	2.232	1.474	1.452	0.840	1.121	0.804	1.380	1.943	0.252	0.699	1.081	1.241	1.090	0.560	1.093	1.030
rmse 변화율	-2.1%	-1.9%	-0.5%	-5.6%	-1.6%	-1.7%	-0.5%	-0.6%	-0.7%	-8.4%	-1.7%	-1.8%	-0.7%	-0.8%	-1.8%	-0.8%	-1.3%

표 2 Optuna 전/후 rmse 비교

나. LightGBM

LightGBM(이하 LGBM)은 많은 예측 대회에서 좋은 성능을 보이는 트리계열의 모델이다. 트리모델에서 예측해야하는 범주의 개수가 많은 경우, 훈련 시간이 오래 소요되고, 과적합될 가능성이 높다. 하지만 LGBM은 비대칭적 트리라는 특성 덕분에 빠른 학습이 가능하며, 분석에 사용한 데이터의 양이 많아 과적합 문제 역시 해소되었다. RandomForest와 마찬가지로 Optuna를 통해 하이퍼파라미터 튜닝을 진행하였다.

다. XGBoost

XGBoost 또한 예측대회에서 인기가 많은 트리모델이며, 과적합 규제기능을 바탕으로 뛰어난 예측 성능을 보인다. XGB도 앞선 트리모델과 마찬가지로 다중공선성이 있는 변수나 영향을 크게 미치지 않는 변수에 대해서는 스스로 중요도를 낮추기 때문에 모든 변수를 사용하였다.

일반적으로 종속변수의 분류 범주가 많으면, 일부 범주에 속하는 데이터가 매우 적기 때문에 실질적인 범주로서의 역할을 하지 못하는 경우가 생긴다. 이를 해결하기 위해 시도별, 성별로 심뇌혈관질환 발생빈도에 대해 구간을 나누었다. 예를 들어, 경기도 남성 데이터의 경우 0-3, 4-5, 6-13 총 3개의 범주로 축소하고, 각 범주에 대한 가중평균으로 대표값을 설정했다.

하이퍼파라미터는 GridSearch와 RandomSearch를 통해 튜닝하였다.

3. 딥러닝 예측모델: LSTM

RNN계열 딥러닝 모델인 LSTM은 단기 데이터와 장기 데이터 모두 기억한 상태에서 학습을 진행하므로 시계열 예측에 적합한 모델이다.

시도별 성별 심뇌혈관질환 발생빈도에 대해 시계열 분해를 진행했을 때, 경기도를 제외한 모든 시도에서 계절성, 추세성 등을 확인할 수 없었다. 즉, 시계열적 특성이 없는 것을 보였고, 각 시도, 성별에 대한 최적 모델에 비해 평균적으로 rmse가 6.69%정도 높았다. 따라서 최종적인 모델로 채택되지 못했다.

4. 분석결과

가. 시도별 성별 최적 모델

각 모델의 2012년~2015년 5-fold CV 결과는 다음과 같으며, 시도별 성별로 rmse가 가장 작은 모델을 선택하였다.

<남성>

시도	강원	경기	경남	경북	광주	대구	대전	부산	서울	세종	울산	인천	전남	전북	제주	충남	충북
RF	1.080	2.189	1.467	1.372	0.827	1.101	0.800	1.372	1.956	0.253	0.687	1.062	1.232	1.081	0.550	1.102	1.017
LGBM	1.084	2.175	1.463	1.362	0.822	1.103	0.801	1.365	1.946	0.252	0.686	1.061	1.224	1.075	0.552	1.089	1.019
LSTM	1.182	2.328	1.580	1.513	0.886	1.178	0.815	1.488	2.133	0.250	0.719	1.094	1.280	1.129	0.565	1.172	1.074
XGB	2.655	2.651	2.093	2.222	3.013	2.590	3.009	2.112	2.018	3.471	3.288	2.828	2.435	2.667	3.353	2.618	2.691
LM	1.269	2.279	1.519	1.467	0.866	1.193	0.805	1.387	2.027	0.273	0.717	1.054	1.227	1.050	0.566	1.062	1.117

<여성>

시도	강원	경기	경남	경북	광주	대구	대전	부산	서울	세종	울산	인천	전남	전북	제주	충남	충북
RF	1.014	2.222	1.506	1.452	0.910	1.086	0.774	1.341	1.878	0.228	0.674	1.028	1.320	1.247	0.527	1.031	0.995
LGBM	1.015	2.215	1.501	1.446	0.909	1.079	0.773	1.335	1.876	0.226	0.671	1.025	1.335	1.245	0.526	1.030	0.993
LSTM	1.045	2.392	1.632	1.571	0.959	1.116	0.795	1.381	2.045	0.248	0.708	1.045	1.444	1.297	0.582	1.062	1.042
XGB	2.712	2.620	2.121	2.128	2.879	2.595	3.043	2.437	1.989	3.471	3.273	2.862	2.235	2.474	3.350	2.681	2.697
LM	1.015	2.453	1.604	1.476	0.954	1.165	0.775	1.374	1.984	0.270	0.683	1.062	1.382	1.302	0.580	1.063	1.101

표 3 전체모델 5-fold CV 결과

전반적으로 LGBM이 가장 우수한 성능을 보였으며, Optuna로 튜닝한 RandomForest도 좋은 성능을 보였다.

나. 결과 해석

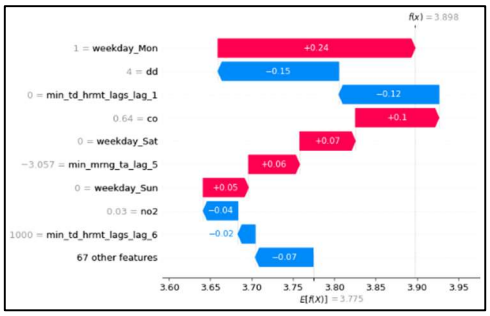

그림 2 개별 예측치에 대한 해석

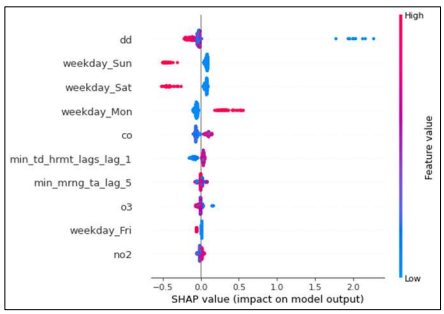

그림 3 각 변수들의 값에 따른 예측치 변동 분포

앞서 선택된 최적의 모델에 대해 Explainable AI 모델인 SHAP을 적용함으로써 독립변수의 기여도

를 바탕으로 예측모델의 결과값을 설명할 수 있다. 예를 들어, 2016년 1월 9일 서울에서 심뇌혈관질환으로 병원을 찾는 남성의 수가 3.1명으로 예측된다면, 그 이유는 토요일이며 월초가 아니고, CO=0.568이며 나머지 변수들은 크게 영향을 미치지 않았기 때문이라고 해석할 수 있다. 즉 SHAP을 통해 익일의 기상예보에 따라 심뇌혈관질환 발생빈도를 예측하고, 이에 더해 발생빈도가 높거나 낮은 이유까지도 설명이 가능해진다.

또한, 예측 모델이 각 변수들의 값을 결과값에 어떻게 반영했는지 해석할 수 있다. 예를 들어 일변수(dd)의 경우 극 월초일 경우 발생률을 높게 예측하고, 월말로 갈수록 낮게 예측하는 경향을 보이며, 주말에는 낮게, 월요일에는 높게 예측하고, CO(일산화탄소) 농도가 높으면 발병률을 높게 예측한다는 것을 알 수 있다.

V. 결과 활용 방안 및 기대 효과

1. 활용 방안 및 기대 효과

현재 기상청 날씨누리에서는 기상 조건에 따른 감기, 뇌졸중 등의 발생 가능 정도를 알려주고 있다. 해당 페이지에 심뇌혈관질환 예상발생빈도를 함께 제공하는 것을 제안한다. 또한, 각 예측값에 가장 크게 기여한 상위 요인들과 대응 요령 등을 안내함으로써 이용자들이 당일 위험 요인을 인지하고, 각별히 조심할 수 있도록 한다. 현재 운영 중인 사이트에 페이지를 추가하는 것이므로 신규 구축에 비해 비용이 적게 발생할 것으로 예상된다.

더불어 하단에 질병관리청 국가건강정보포털 심뇌혈관질환정보 페이지 링크를 삽입하는 것도 가능하다. 해당 페이지에는 심뇌혈관질환의 종류부터 그 증상, 원인 및 예방에 대한 정보가 있어 추가적인 도움을 줄 것으로 기대된다. 또한 중앙-권역-지역 심뇌혈관질환센터 목록을 제공해 병원에 대한 접근성을 높일 수 있다. 이는 고위험군에게 조기검진을 유도하여 중증질환으로 발전하는 것을 예방할 수 있다.

심뇌혈관질환의 원인은 보통 나이나 생활습관으로 알려져 있지만, 심뇌혈관질환 발생 예측모델을 통해 다양한 정보를 제공함으로써 심뇌혈관질환 발생에서 기상요인에 대한 인식을 제고할 수 있다. 특히 노년층은 심뇌혈관질환에 취약한데, 전연령대 중 가장 높은 TV시청률을 보이는 특징이 있다. 따라서 해당 결과와 해석을 바탕으로 각 방송국 기상뉴스 보도국과 긴밀히 협력하여 적극적으로 시민들에게 알리는 것이 주효할 것으로 보인다. 앞서 제시한 활용방안을 통해 궁극적으로 국민 건강 증진에 기여할 수 있을 것으로 기대된다.

VI. 참고문헌

[i] 「제1차 김뇌혈관질환관리 종합계획[2018~2022」, 보건복지부, 2018.09.04.

[ii] 하경화, 김창수, 서민아, 강대용, 김현창 and 신동천. (2011). 「미세먼지 농도와 심뇌혈관계 질환으로 인한 사망과의 관련성」. 『Clinical Hypertension』, 17(2), 74-83. X

[iii] Yoneyama, K., Nakai, M., Higuma, T. et al. 「Weather temperature and the incidence of hospitalization for cardiovascular diseases in an aging society」. 『Sci Rep 11』, 10863 (2021).,
Gonçalves FL, Braun S, Dias PL, Sharovsky R. 「Influences of the weather and air pollutants on cardiovascular disease in the metropolitan area of São Paulo」. 『Environ Res』. 2007 Jun;104(2):275-81.,
Bijelović S, Dragić N, Bijelović M, Kovačević M, Jevtić M, Ninkovic Mrđenovački O. 「Impact of climate conditions on hospital admissions for subcategories of cardiovascular diseases」. 『Med Pr』. 2017 Mar 24;68(2):189-197.,
하경화, 김창수, 서민아, 강대용, 김현창 and 신동천. (2011). 「미세먼지 농도와 심뇌혈관계 질환으로 인한 사망과의 관련성」. 『Clinical Hypertension』,.17(2), 74-83.,
Lee, S., & Yeo, I-K. (2020). 「Predicting the number of disease occurrence using recurrent neural network」. 『The Korean Journal of Applied Statistics』. 33(5), 627-637.

[iv] 3과 동일

[v] 김희경, 강인경, 이재원, 이영섭, 「연속적 결측이 존재하는 기온 자료에 대한 결측복원 기법의 비교」, 『응용통계연구』.29(3),549-557., 2016.

[vi] 대한민국 기상청, 「[기상기후 빅데이터, 날씨마루] 5강 이상치 결측치 확인 및 처리」,2018.12.18

[vii] Lee, S., & Yeo, I-K. (2020). 「Comparison of forecasting models of disease occurrence due to the weather in elderly patients」. 『The Korean Journal of Applied Statistics』. 29(1), 145-155. ,
Lee, S., & Yeo, I-K. (2020). 「Predicting the number of disease occurrence using recurrent neural network」. 『The Korean Journal of Applied Statistics』. 33(5), 627-637.

[viii] Abrignani MG, Lombardo A, Braschi A, Renda N, Abrignani V. 「Climatic influences on cardiovascular diseases」. 『World J Cardiol』. 2022 Mar 26;14(3):152-169.