기상청

기상에 따른 혈관질환 발생 예측 모델 개발

참 가 번 호	220081	팀 명 ※ 반드시 참가신청 시 작성한 팀명	선크림

1. 분석 데이터 설명

1-1 사용 데이터 설명

1) 혈관질환 발생에 미치는 기상조건

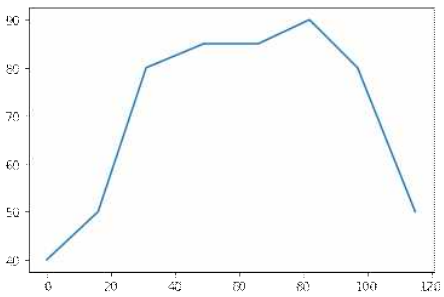
- 기상 변화를 대표할 수 있는 기온, 일교차, 습도 등 12개의 기상 변수 사용
- 기상조건 변화는 당일 질병 발생률보다 1~2일 후 질병 발생률에 영향을 주는 것이 일반적이므로 독립변수와 종속변수 간 이틀의 시간차를 두고 모델 개발을 진행하고자 했으나, 예보데이터 내 결측값 및 누락값으로 인해 동시간대로 결합 및 분석 진행

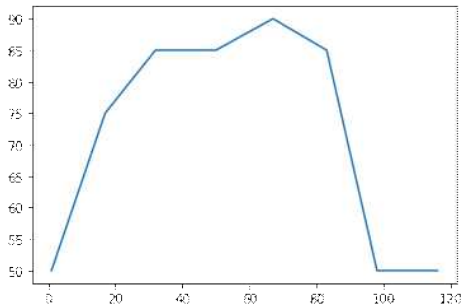
2) 외부에 노출되기 쉬운 시간대

- 제공 변수 내 관측 시점이 오전 2시부터 오후 23시까지 3시간 단위로, 총 8개의 시점 존재
- 이 중 비교적 기상 수치가 두드러지는 오전 8시와 오후 17시 사이, 총 4개의 시점만 사용
- 오전 8시와 오후 17시는 사람들이 외부에서 활발하게 활동하는 시간대이므로, 비교적 날씨와 건강 간 영향이 크다고 예상

[시간대별 기상 수치 대표 군집]

- 예보데이터
- 관측 시간대: 2시, 5시, 8시, 11시, 14시, 17시, 20시, 23시
- x축: 관측 시간대, y축: 기상 수치





1-2 제공 변수 및 추가 변수

①. 사용한 변수 리스트 (변수명, 출처, 간단한 설명, 기간)
②. 사용한 모든 데이터의 기간은 2012년 ~ 2016년으로 동일
③. 제공 변수 내 'hour' 칼럼값 0800 ~ 1700만을 사용
④. 추가 변수의 기상관측 단위는 일 단위 (관측된 시간대를 평균으로 결합하여 제공됨)
⑤. 추가 변수의 황사관측 단위는 시간 단위

⑥. '일교차' 변수의 경우, 파생 변수로 기상관측 데이터 내 '최고기온'에서 '최저기온'을 뺀 값

⑦. 'area' 변수의 경우, 각 지역별 더미변수로 변환하여 사용 → 'area_(지역명)'

구분	예보 데이터 및 백병원 데이터	구분	백병원 및 미세먼지/기상관측 데이터	
제공 데이터 (출처: 기상청)	yyyymmdd		sex & frequency	area
	hour & forecast	추가 데이터 (출처: 기상자료개방포털)	1시간평균 미세먼지농도(㎍/㎥)	평균 지면온도 (°C)
	강수량		평균기온(°C)	일교차 (최고기온 - 최저기온) (°C)
	적설		최저기온(°C)	평균 이슬점온도(°C)
	습도		최고기온(°C)	평균 상대습도(%)
	하늘상태		평균 풍속(m/s)	평균 증기압(hPa)

2) 기온과 혈관질환 간 관계

- 기온이 낮아지면 혈관이 수축되어 혈압은 높아지고 체내 혈액의 점성은 증가
- 마찰력이 커져 혈액의 흐름이 둔화되어 혈전이 생성될 위험이 높아짐
- 혈관이 막히거나 터지는 원인이 되어 고혈압과 함께 뇌졸중을 일으키는 원인으로 작용함
- 기온이 높아지면, 특히 폭염으로 땀을 많이 흘리면, 혈액의 농도가 짙어져 생긴 혈전으로 관상동맥 혈관이 막혀 심근경색증이 발생

3) 미세먼지와 혈관질환 간 관계

- 대기 중 단일오염 물질이 뇌혈관 질환 사망자에 유의한 영향을 준다는 연구 결과 존재
- 따라서 '1시간 평균 미세먼지 농도'를 독립변수로 추가

4) 기상관측 데이터

- 결측치 및 누락값을 대체하기 위한 기상관측 데이터 사용
- 예보 데이터 내 풍속, 풍향, 기온, 강수형태, 강수확률 변수를 기상관측 데이터 내 평균 풍속, 평균기온, 최저기온, 최고기온 등의 변수로 대체
- 기상관측 데이터의 경우, 일 단위로 관측되었으며 각 지역 내 행정구역에 대한 시간대별 관측값을 모두 평균 내어 제공됨

2. 분석 프로세스

2-1 분석 프로세스

2-2 분석 과정

1) 데이터 수집 및 전처리

①. 추가 데이터 수집 및 결합

- 관측 빈도가 다른 두 변수(강수량, 적설)가 존재하는 경우, 두 칼럼에 대한 평균값으로 통합
- 제공 데이터와 추가 변수 간 inner join을 활용한 데이터 결합

②. 이상치 및 결측치 처리

- 예보 데이터의 경우, 결측치 처리에 필요한 충분한 데이터가 부재함에 따라 2012.01.01. ~ 2015.12.31. 기간 중 결측치가 있는 행 제거
- 기상관측데이터의 경우, 이상치 범위를 25% ~ 75%로 설정
- 기상관측데이터인 경우, 결측치를 spline메소드를 이용하여 보간

③. 기존 변수 대체

- 방대한 양의 결측값 및 누락된 값이 존재하는 제공된 데이터 내 변수 중 기상관측 정보로 대체 가능한 변수(풍속, 풍향, 일최고기온, 일최저기온 등) 선정 및 대체

④. 다중공선성 확인 및 회귀분석 결과

- 독립변수들 간 다중공선성 확인 후 0.9 이상인 두 변수(평균 지면온도, 평균기온)에 대해 하나의 변수(평균기온)로 축소
- 학습 및 테스트 데이터 내 회귀분석 결과, p-value가 0.61 이상(강수량, 적설)인 변수 제거
- 정확한 데이터 부재(대량의 누락 및 결측값)로 혈관질환 환자 수를 예측하는데 유의미하지 않다고 판단하는 기준이 되는 p-value를 다소 높게 설정

⑤. 정규화

- 모든 데이터 포인트를 동일한 정도의 스케일(중요도)로 반영
- Min-Max Scaler (최소-최대 정규화) 사용

2) 데이터 분리

- 학습 데이터: 2012년 2월 ~ 2015년 12월 (검증 데이터로 사용된 기간 내 데이터 제외)
- 테스트 데이터: 2012년 3월, 2013년 6월, 2014년 9월, 2015년 12월 (계절 별 한 달치)
- 검증 데이터: 2016년 1월 ~ 2016년 12월

[최종 학습 변수]

구분	예보 데이터 및 백병원 데이터	구분	미세먼지 및 기상관측 데이터
제공 데이터 (출처: 기상청)	yyyymmdd	추가 데이터 (출처: 기상자료개방포털)	1시간평균 미세먼지농도(㎍/㎥)
	hour & forecast		평균기온(°C) & 일교차 (최고기온 - 최저기온) (°C)
	습도		평균 풍속(m/s)
	하늘상태		평균 이슬점온도(°C)
	sex & frequency		평균 상대습도(%)
	area		평균 증기압(hPa)

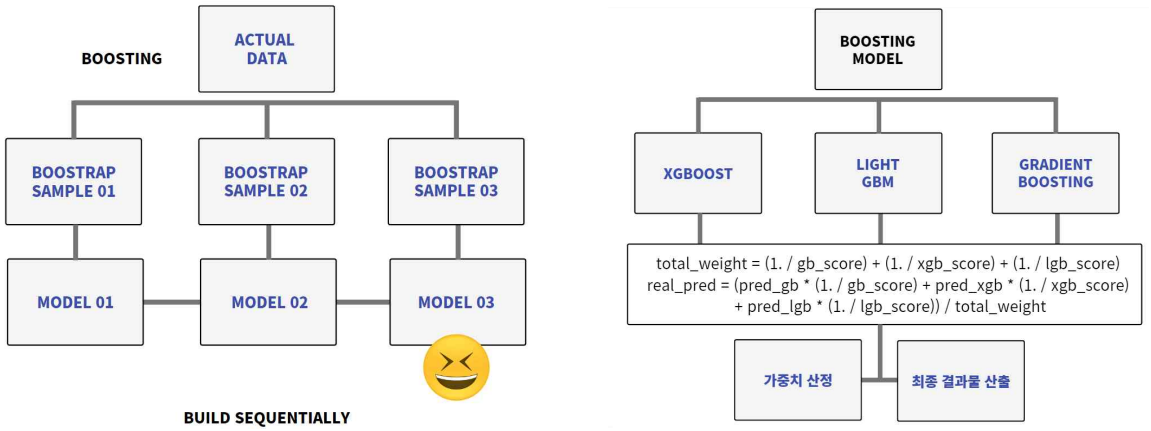
4) 회귀 모델

①. 활용 모델

앙상블	부스팅	XGB Regression
		GradientBoosting Regression
		LightGBM Regression

②. 3가지 앙상블을 이용한 최적의 결과 도출

- 총합이 1이 되도록 모델별 weight를 결정
- train set의 결과인 RMSE의 역수에 비례하여 가중치 산정
- 각 모델의 가중치와 예측값을 곱한 최종 결과물 산출



5) Time Series Split

- 시계열 분할 교차검증은 그 이름 대로 시계열 데이터를 사용한 분할법
- 미래와 과거의 데이터를 혼동한 훈련 세트로 학습을 실시하여, 모델이 각 시점에 어떤 일이 일어났는지 제대로 학습할 수 없음
- 고정된 시간 간격의 시계열 데이터 교차검증에 활용

3. 분석 결과

3-1 3가지 모델 결과 작성

활용모델 및 검증값	모델 성능 평가결과 (RMSE)	2016 검증값 순위 (2022.08.08, 14: 40 기준)
XGBoost	1.3648	18위
GradientBoosting	1.3834	18위
LightGBM	1.3659	13위

```
total_weight = (1. / gb_score) + (1. / xgb_score) + (1. / lgb_score)
real_pred = (pred_gb * (1. / gb_score) + pred_xgb * (1. / xgb_score)
+ pred_lgb * (1. / lgb_score)) / total_weight
```

최종 모델 및 검증값	2016 검증값 순위 (2022.08.08, 14: 40 기준)
3가지 앙상블	12위

3-2 최종 선정 모델

- 3가지 앙상블을 이용한 최적의 결과 도출

1) 파라미터

최종 모델(XGBoost) 사용 파라미터			
learning_rate	0.1	n_estimators	20000
max_depth	3	subsample	0.7

최종 모델(GradientBoosting) 사용 파라미터			
subsample	0.70	min_samples_leaf	5
n_estimators	1000	min_samples_split	2
max_depth	3	learning_rate	0.1

최종 모델(LightGBM) 사용 파라미터			
bagging_fraction	0.72	scale_pos_weight	1.5
num_iterations	20000	lambda_l1	0.1
max_depth	7	lambda_l2	0.35
min_data_in_leaf	8	early_stopping_rounds	300
learning_rate	0.01	n_splits	4
colsample_bytree	0.72	random_state	42

2) RMSE

- 최종 선정 모델은 3가지 앙상블 회귀 모델이며, 최종 순위는 12위 (2022.08.08., 14:40 기준)

4. 활용 방안 및 한계점

4-1 활용 방안

1) 경고 시스템 구축

①. 혈관질환에 영향을 미치는 기상 조건 변화를 빠르게 감지할 수 있는 시스템 구축

②. 혈관질환 발생 조기 예방을 위해 기상 감지 시스템 간 위험 단계를 나누어 경고 시스템 구축

2) 유연한 지역 쉼터 운영

①. 날씨 예보에 따른 기온 변화에 따라 지역 내 쉼터를 유연하게 운영하는 방침 재정

②. 기상 조건 중 하나인 '일교차'와 같이 혈관질환에 가장 큰 영향을 미치는 '기온' 변화에 따른 여름철 무더위쉼터 및 겨울철 한파쉼터 운영

3) 병원 내 혈관질환 환자 발생 대비 시스템 구축

①. 실시간 기상 조건 변화 정보를 기반으로 발생 가능한 환자 수를 예측

②. 예측한 환자의 수가 많은 경우, 이를 대비하여 응급 혈관질환 환자 수용 및 응급처치를 위한 대책 및 환경 마련

③.혈관질환 환자에 대해 예방관리 교육 및 건강검진 실행

4-2 한계점

1) 기상 변화와 혈관질환 발생 간 인과관계

- 당일 기상 변화가 당일 혈관질환 발생에 영향을 미치기보다는 당일 혈관질환 발생에 대해 하루 ~ 이틀 전 기상 변화가 영향을 미침
- 그러나 한정된 제공 데이터 및 다량의 결측 및 누락값으로 인해 독립변수와 종속변수 간 시차 없이 모델 개발
- 검증 결과뿐만 아니라 모델 자체 RMSE(정확도)에 중점을 두어, '기상에 따른 혈관질환 발생 예측 모델' 개발
- 향후 정확하게 관측된 기상 데이터를 기반으로, 독립변수와 종속변수 간 시차를 이용해 최적 모델 개발 가능

2) 혈관질환 환자(백병원) 정보

- 환자에 대한 '성별'과 총 환자 '발생 수' 정보만 주어져 있어 기존에 환자의 건강상태가 혈관질환 발생률에 미치는 영향을 고려할 수 없음
- 향후 연령, 세부적인 거주지(행정동), 고혈압 및 당뇨병 유무 등 세부 정보를 기반으로 비교적 정확도 높은 모델 개발 가능

기상청

[참고 문헌]

온도변화와 허혈성 심장질환 및 뇌혈관 질환에 따른 사망자수와의 관계연구, 김양희, 2015.08

서울시 환경요인과 월별 뇌혈관 질환 사망자의 연관성, 박희원, 2009.11