데이터 셋 관련 질문 :

test data set 이랑 predict data set 의 size 가 다른데 왜 그렇게 모델이 구성된 것인지?

:

각종 데이터 포털과, 시군구별 지자체에 문의하여 사망자수 데이터를 확보하려고 했으나, 공식적으로 취합된 데이터라는 것이 존재하지 않아서, 시도별과 경기도 시군구별로 40여개의 지역에 대한 데이터를 수집할 수 있었다.

사망자수를 예측할 때, test set data 와 동일하게 40여개의 지역에 대해서만 예측결과를 구한다는 것은 병상 부족 지역을 선정한다는 목표에 너무 광범위한 지역 단위이기에 유의미한 결론을 도출하기에 한계가 있다고 판단했다. 따라서

예측 시에는 모든 시군구별에 대하여 예측하였다.

추후에 다른 시도에 대해서도 시군구별 사망자수 데이터를 확보하여, 현재의 모델을 수정한다면 지금보다도 훨씬 성능이 좋은 모델로 개선할 수 있을 것이라고 기대하고있다.

선정 기준에 대한 질문:

1. 사망자수는 누적된 값이고, 병상 수는 아닌 것 같은데, 병상수/사망자 수로만 기준을 설정하게 된 이유는 ?

: 코로나19는 치명률이 그리 높지 않은 감염병인 만큼 일일 사망자의 수가 크지 않고, 그렇기에 일일 사망자수로는 유의미한 결론을 도출하기 힘들다. 본 프로젝트의 목적은 병상수가 골고루 분포되었는지를 확인하고, 병상수가 부족하다고 판단되는 지역을 선정하는 것이다. 그렇기에 병상수가 부족하다고 판단할 수 있는 기준이 필요한데, 필요 병상수에 직접적인 영향을 미치는 변수인 고위험 확진자군의 수를 대변할 수 있는 변수로 누적 사망자수를 설정하였고, 이를 전지역에 적용하여 구한 사망자수 대비 병상수의 비로 병상수의 분포를 확인하기에 충분하다고 판단하였다.

1. 병상 수로 그룹화한 이유는 ?

: 병상 수로 그룹화한 이유는, 선정된 지역에 추가할 병상의 수를 추후에 결정할 때 해당 지역의 현재 의료자원 및 인력의 규모를 고려해야하기 때문이다. 각 지역의 병원별 의료 인력 및 자원의 분포를 전부 취합할 수 없으므로, 그에 대응하는 지표로 병상 수 자체를 의료자원의 규모로 설정하였고 따라서 병상 수로 그룹화하여 해당 지역이 현재 어느정도 규모의 의료자원이 할당된 지역인지 파악될 수 있게하고자 하였다.

1. 병상 수 확보의 기준을 사망자수로 잡은 이유?

: 코로나19 확진자수가 증가하면서 병상이 모든 환자들을 수용하기에 더더욱 부족해졌다. 따라서 확진자 중 악화된 증세를 보이는 이들을 중심으로 병상을 할당해야할텐데, 위중증 환자수 데이터는 따로 구할 수 없었고, 확진자 중 사망으로 이어지는 환자들이 현재 확보할 수 있는 데이터 중 병상이 필요한 고위험군으로 판단할 수 있는 합리적인 지표라고 생각하였다.

1. 실제 사망자 수를 그대로 사용하는 것이 아닌 예측된 사망자 수를 사용 하였는지

: 변수 중요도에 따른 결과값을 평균으로 나타낸 회귀 모델을 만들어냈는데, 가중치가 부여된 각 관측값(독립변수)에 따른 적합값으로 평균적으로 얼마가 나와야하는지 파악하기 위해 회귀모델을 만듦

독립변수 관련 질문 :

* 독립변수에 요양원수를 넣은 이유?
  + 고령인구가 바이러스에 대한 치사율이 더 높기 때문에, 고령인구가 밀집된 요양원도 고려해야 한다고 생각했다.
* 고령인구 나이를 5세 단위씩 세분화해서 넣은 이유?
  + 고령인구 중에서도 나이대별로 치사율이 다르며, 나이가 많을수록 치사율이 급격히 증가하기 때문에 세분화했다.

머신러닝 과정 질문 :

* 왜 MinMax scaler를 사용했는지?
  + ​​모든 값의 범위를 0에서 1 사이로 변환하면서도, \*\*분포가 유지된다는 특성이 있기 때문
  + \*\* 분포가 유지되어야 하는 이유? - 말 그대로 인구수는 병상수 예측에 중요한 변수이기 때문에 해당 데이터의 특성을 최대한 살려야 된다고 생각했기 때문
  + 정규분포(bell-shape)가 아니거나 표준 편차가 작을 때 효과적
* 이상치를 다 제거 안 한 이유 or 하나만 제거한 이유 ?

: 학습 데이터 셋의 수가 적은편이라 최대한 데이터 수를 살리기 위함. 또한 ‘인구수'라는 데이터의 특성 상 지역 별로 차이가 크다는 특성을 가지고 있어서 이상치 또한 포함시켜서 지역별 특성을 반영해야 한다고 생각했음. 따라서 다른 지역들보다 이상치가 월등히 커서 모델 형성에 영향을 크게 미치는 서울을 제외하고는 다 포함시켰음

* 평가 지표로 왜 rmse를 사용했는가? (근데 다른 지표도 랜포가 제일 낮았어서 이 질문은 걱정 ㄴ)

: RMSE는 지표 자체가 직관적이며 예측변수와 단위가 같으며, 잔차를 제곱하기 때문에 다른 "큰 오류값 차이에 대해서 크게 패널티를 주는" 이점이 있다.

* 각 지역별로 인구수가 다르기 때문에, 인구수가 많은 지역에만 사망자가 몰린다는 단순한 결론에 내릴 수 있지 않을까요?
  + > 그래서 그룹화 한거 아녀..?
  + > 병상 수로 그룹화해서 탑3를 가져온 것이 인구 수에 많이 의존한 것인지, 어떤 요인을 기준으로 나눈 건지 보여줘야할 것 같음
  + 병상수로 그룹화 해도 그룹 내에서

어떤 요인을 기준으로..?

\* 격리병상수와 남은 격리병상수는 사망자수와 선형관계를 나타내지만 인과관계를 나타내지 못함

\* 우리가 사용하는 예측 모델에 포함되지 않는 어떤 변수가 반응변수(response ariable)와 하나 이상의 예측변수(predictor variable) 모두에 영향을 줄 때, 이것을 혼선자(confounder)로 다룹니다.

\* 혼선(confounding)은 어떤 변수들이 다른 변수들에 영향을 주고 있는지 판단하기 어렵게 만들지만, 항상 예측 작업을 더 어렵게 만들지는 않습니다.