1p

안녕하세요, 지금부터 STAT-IT 팀 발표 시작하겠습니다.

저희는 병상 수 중심으로 코로나 현황을 파악하고 병상이 부족할 예상 지역을 선정하는 프로젝트를 진행했습니다.

4p

먼저 프로젝트를 간단하게 소개하겠습니다.

지난 수개월동안 확진자 및 위중증 환자 수가 증가하며 병상 부족 문제가 심각해졌고, 정부는 이에 따라 격리병상 확보 정책을 세웠습니다. 병상 가동률은 안정되는듯 보였지만, 하지만, 현실은 병상 부족으로 사망자가 생겨나는 상황이었습니다.

따라서 병상 확보의 적절성을 판단하고, 병상이 부족한 지역에 대한 대책 필요함에 따라 프로젝트를 기획하게 되었습니다.

5p

모델링에 앞서 코로나 사망자 및 격리병상의 현황을 시각화를 통해 분석했습니다.

우선 사망자의 90% 이상이 60세 이상의 고령자인 것을 파악했고, 고령환자 일수록 치명률이 급격하게 증가하는 것을 확인했습니다.

6p

시도별 누적 사망자 분포는 수도권 지역이 다른 지역들에 비해 월등히 높은 수치를 띄고 있음을 확인 할 수 있습니다.

7p

또한 지도로 시각화 했을 때, 인구 수가 많은 수도권에서 상대적으로 병상이 부족한 것을 확인했습니다.

8p

이와 같은 현황 분석을 통해 전국적으로 병상 수가 인구 수, 고령인구 수, 사망자 수 대비 적절하게 확보되어 있지 못하다는 것을 확인 할 수 있었습니다.

'적절한' 병상 수가 '미리' 확보될 수 있게끔 머신러닝 모델을 통해 사망자수를 예측하고, 현재 확보된 병상 수와 비교하여 병상이 더 필요할 지역을 선정하는 것을 궁극적인 목표로 설정하였습니다.

9p

데이터 수집 및 모델링 과정은 다음과 같습니다.

11p

패스

12p

병상 및 병원 수 관련 데이터들을 수집하기 위해 크롤링 하였고, 다음은 크롤링에 사용된 코드입니다.

13p

말 안하고 패스

14p

다음으로 지역 선정 과정 및 결과에 대해 말씀드리겠습니다.

15p

지역 선정을 위해 다음과 같은 5가지 모델을 사용하였습니다. 모델 선정 기준으로는 RMSE를 선택하였고, 가장 작은 값을 가진 모델인 RandomForestRegressor를 최적 모델로 선정하였습니다.

16p

독립변수와 종속변수의 설정은 다음과 같습니다.

17p

또한, 대부분의 변수가 양의 선형관계를 띠는 것을 확인했습니다.

<랜덤 포레스트 부분>

18p

데이터 전처리 과정 첫 번째로, 이상치를 제거하였습니다. 가장 큰 이상치를 가지는 서울특별시의 데이터를 제거하여 분포를 크게 줄였습니다.

19p

그 후, 다른 변수들에 비해 값의 단위가 큰 인구수 데이터들을 MinMaxScaler를 이용하여 스케일링 하였습니다. 보시다시피 값의 범위가 0과 1 사이로 축소 되었지만, 분포는 동일한 것을 확인할 수 있습니다.

20p

이렇게 전처리 과정을 마친 후, 데이터셋을 훈련시켜 데이터셋을 train, valid, test 셋으로 나눕니다. 먼저, train 세트에 n\_estimators를 500으로 지정하여 변수 중요도 및 예측 오차를 확인하였습니다.

21p

머신러닝 모델 결과 중요 변수인 60대 인구수, 병상 수 등을 추출하여 고위험구역 현황을 파악할 수 있었습니다.

22p

이어서 Grid search를 이용해 random forest의 주요 파라미터를 튜닝합니다. 시행 결과, ‘max\_features’가 7, n\_estimator가 1700인 모델을 최종 모델로 선정하였습니다.

그 후, test 세트에 적용시켜 최종 예측 성능과 오차를 확인하였습니다.

23p

마지막으로, 실제 데이터에 모델을 적용하여 시군구별 사망자 수를 예측하였습니다.

24p

위 그림들은 하나의 decision tree를 시각화한 것입니다.

25p

사망자수 예측결과는 다음과 같습니다. 사망자수가 100명이하인 지역은 106개, 200명 이하는 58개, 최대값인 491명 이하는 42개 정도의 분포를 띄고 있습니다.

26p

다음으로는 지역 선정 과정을 말씀드리겠습니다. 우선 병상수의 규모에 따라 지역을 선정하기 위해, 다음과 같이 5개의 그룹으로 나누었고 지역을 선정하기 위한 기준 수치는 사망자수 대비 병상수로 지정하였습니다. 그룹 0은 병상이 '0' 인 지역으로 의료 시설 추가 확보가 단기에 이루어지기 어려운 지역으로 판별하여 지역 선정에서 제외하였습니다.

27p

그룹 1의 예측 사망자수 대비 병상수 비의 분포이며, 지도에선 짙은 붉은 색을 띌수록 작은 값을 의미합니다. 우측은 하위 3개의 지역입니다. 이와 같은 방식으로 각 그룹별 예측 사망자수 대비 병상수 비의 분포를 확인하겠습니다.

28p

다음으로는 그룹 2의 예측 사망자수 대비 병상수 비의 분포입니다.

29p

다음으로는 그룹 3의 예측 사망자수 대비 병상수 비의 분포입니다.

30p

마지막으로 그룹 4의 예측 사망자수 대비 병상수 비의 분포입니다.

31p

최종적으로 각 그룹에서 3 개의 지역씩, 총 12개 지역을 병상 추가 확보가 가장 필요한 지역으로 선정하였습니다.

32p

마지막으로 본 프로젝트의 활용 방안 및 기대효과에 대해 말씀드리겠습니다.

33p

현재 보건복지부에서는 '병상가동률' 이라는 지표를 통해, 병상이 부족한지에 대해 판단하며 올해 초 병상 가동률이 안정적이라고 발표했습니다.

34p

그러나 동시에 병상 부족으로 인한 각기 다른 사망 사례를 수십건 확인 할 수 있으며 전문가들도 병상가동률이라는 지표가 병상 부족 여부를 제대로 대변하지 못함을 말하고 있습니다. 따라서 병상 수의 확보를 이와 같은 지표가 아닌, ‘예측 모델’에 근거한다면 정책 시행시 타당성을 확보할 수 있을 것 입니다.

→ 현재 사용하는 병상가동률이란 지표는 실효성이 떨어지는 것으로 판단할 수 있습니다. 이에, ‘예측 모델' 기반 정책 시행을 통해 타당성을 확보할 수 있습니다.

35p

나아가, 머신러닝 모델 예측 결과를 이용하여 사망자 관련 방역 조치가 적절하게 이루어진 것인지 시군구별로 평가하여 보완이 가능할 것이며 추후에 발생 가능한 코로나 변이 바이러스 또는 신종 전이바이러스 창궐시에도 해당 모델을 기반으로한

→ 또한 추후 발생 가능한 신종 바이러스 창궐 시에도 해당 모델을 적용시켜 의료 시설 및 의료진 보완까지 가능할 것입니다.

36p

이상 STAT-IT 팀의 발표를 마칩니다. 감사합니다.

7분 50초