안녕하세요 사고예측 2팀 이재연, 방현성, 박세준입니다. 저희는 교통사고데이터를 기반으로 각각의 도로의 위험지수를 예측하는 모델에 대해 발표하겠습니다.

먼저 개요, 데이터, 예측모델 순으로 발표를 진행하겠습니다.

저희는 예측모델을 만들기 앞서, 현재 세종시의 교통사고 현황이 어떻게 되는지 확인해봤습니다. 그래서 TAAS 세종시 교통사고 데이터에서 가장 최근 데이터인 2022년 1월부터 2022년 12월까지의 데이터를 사용해 EDA를 진행하였습니다.

먼저 세종시 교통사고 데이터분석결과, 사상자는 경상자가 가장 많았고 그 다음 부상자, 중상자, 사망자 순으로 많았습니다.

피해, 가해 운전자 연령대를 살펴보면 피해운전자는 30, 40대가 가장 많았고, 가해 운전자는 40,50대가 많았습니다.

왼쪽그림은 피해자 차량 종류, 오른쪽은 차량종류를 분석한 그림입니다. 두가지 모두 승용차가 대부분을 차지했고, 피해자 데이터에는 자전거와 보행자도 꽤 많은 것을 확인할 수 있었습니다.

기상상태에 따른 사고 현황은 예상대로 눈과 안개가 있을때 사고가 많이 일어났음을 알 수 있었습니다. 오른쪽 사진은 세종시 동, 읍별로 사상자 합계를 나타낸 것입니다. 전동면, 집현동이 가장 많은 사고가 있었음을 알 수 있었습니다.

저희는 이러한 세종시 교통사고에 관한 배경을 바탕으로, 위의 표처럼 시간대가 주간, 소통이 원활, 사망자수가 1, 중상자수가0 , 경상자수가 2. 부상자수가 1. 사상자 수가 5일때, 이 사고가 난 도로의 도로 위험 지수를 예측하는 모델을 만들고자 했습니다.

도로위험지수를 계산하는 식은 TAAS에 있는 도로위험지수 산출식을 사용하였습니다. 사진을 보면 도로위험지수는 10 곱하기 도로위험도 (사고건수와 사상자수와 사상자환산계수를 모두 곱한값) 곱하기 날씨, 교통소통상황, 주야시간대를 곱한값임을 알 수 있습니다.

여기서 저희는 세종시의 도로 형태를 보면, 일반국도가 대다수임을 알 수 있습니다. 그래서 왼쪽 사진을 보면 일반국도에서의 기상에 따른 영향계수가 1임을 알 수 있습니다. 그래서 저희는 기상상태를 1로두고, 소통과 시간대는 각각 원활할때와 주간일때로 두고 계산을 진행하였습니다.

저희는 이러한 계산식을 바탕으로 2019년 6월 부터 2022년 12월 세종시 도로별 교통사고 데이터를 가지고 모델예측을 진행하였습니다. 이 데이터는 3년간 5건이상의 사고가 발생한 도로명과, 사망자수, 경상자수등의 정보를 담고있습니다. 그래서 저희는 이 데이터에 도로명과 사망자수, 중상자수, 부상자수를 이용하여 사상자환산계수, 도로위험도, 사상자수를 추가해주었습니다.

또한 같은 도로명끼리 합쳐 도로 위험지수를 계산하였고, 도로명에 따른 위경도또한 추가해주었습니다.

이 데이터의 위경도를 바탕으로 지도에 시각화한 결과 입니다.

다름은 도로별 도로위험지수를 시각화한 결과입니다. 이때 도로위험지수가 8보다 작으면 안전, 8~20이면 주의, 20~27이면 위험, 27이상이면 심각이라는 tass정보를 사용하였습니다.

이렇게 전처리한 데이터를 가지고 독립변수 x를 사망자수, 경상자수, 부상자수, 사상자환산계수, 사상자수로 두고, 종속변수 y를 도로위험지수라고 두고 linear regression, 릿지, 라쏘, 일렉트릭넷을 사용하여 도로위험지수를 예측하였습니다.

예측에 앞서 각각의 변수들이 모델에 얼마나 기여하는지를 확인해보았습니다. 이때 사망자수, 사상자환산계수, 사상자수가 가장 크게 기여함을 알 수 있었습니다.

다음은 linear regression 모델, lasso 모델, elasticnet모델을 적용한 결과입니다. 각각 훈련데이터와 테스트데이터의 정확도는 선형회귀는 0.60,0.50이고, 라쏘모델은 0.57,0.44 , 일렉트릿넷모델은 0.60,0.49가 나왔습니다.

다음은 릿지모델을 알파값을 다르게 하여 확인해본 결과 입니다. 세개의 결과를 살펴보면 알파값의 따라 큰 차이가 없음을 알 수 있었습니다.

각각의 모델의 예측값과 정답값을 비교해본 표 입니다. 이중 한가지만 살펴보면, 선형회귀와 라쏘 각각 19.876의 정답값을 23.33과 24.36으로 예측한 것을 알 수 있습니다.

다음은 알파가 5일때 릿지 모델과 일레스틱모델입니다. 여기에서도 19.876의 정답값을 22.618, 23.402로 예측한 것을 살펴볼 수 있습니다.

결론… 아쉬운점.. 정확도를 높일수있는방법..?

감사합니다. 질문있으신가요?