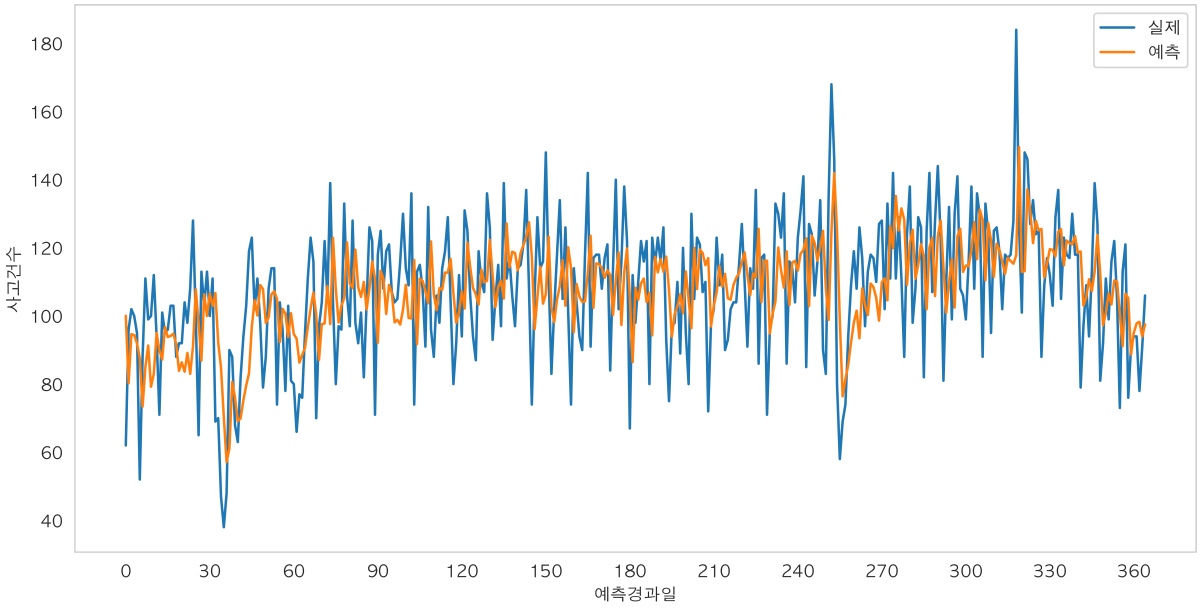
교통사고 유발요인 분석 및 사망사고 예측

- 멀티캠퍼스 프로젝트 2
- 개발 기간
  - 7월 28일~ 8월 18일
- 개발 인원
  - 5명
- 본인 역할
  - LSTM : 서울시 시계열 사고건수 예측
  - Imbalance Classification: 전체사고중 사망사고 분류 예측

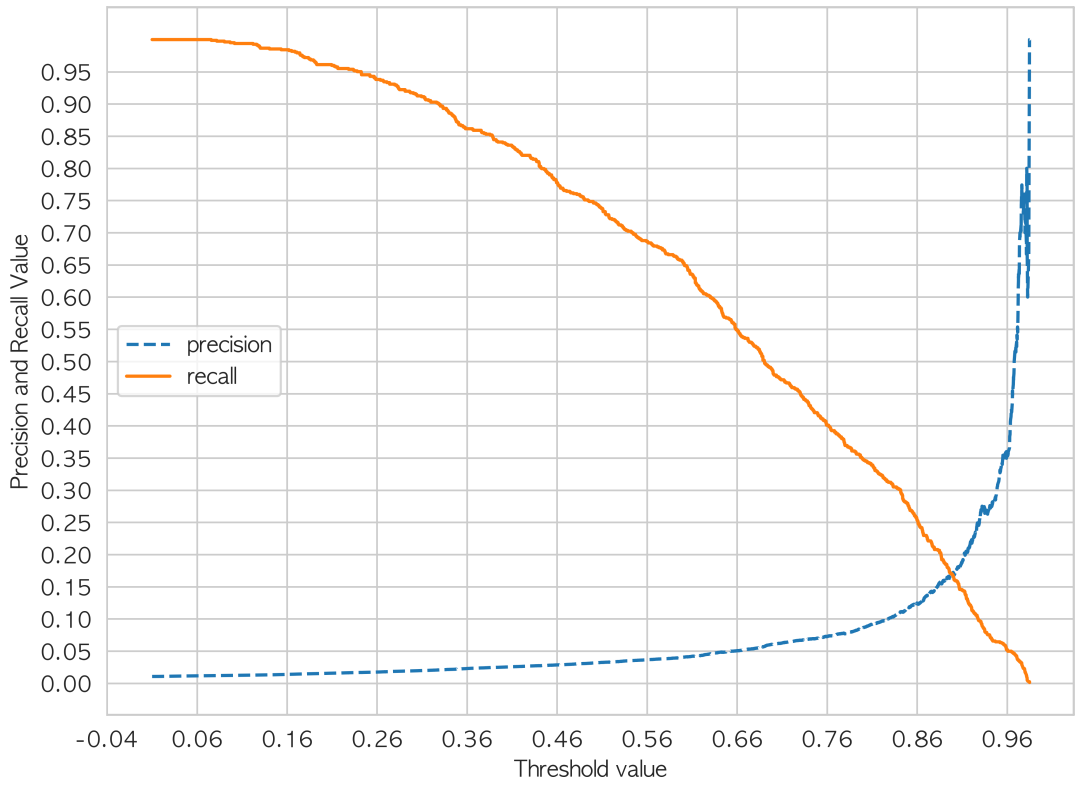
분석

- 대한민국의 교통사고 사망률은 OECD 주요국 대비 2위, 인구 1천명당 사고건수는 OECD 13개국 대비 1위를 차지하고 있었습니다. 저희는 사고 발생시 사망에 이르는 요인을 분석하여, 사망률을 감소시킬 수 있도록 해결방안을 제시하는 것을 목표로 하였습니다
- 사망률에 대한 EDA 분석
  - 기후 : 안개 낀 상황에서 교통사고 사망률이 가장 높았습니다
  - 법규: 국내 전체 사망률은 꾸준히 감소하나 과속으로 인한 사망률이 11퍼센트로 가장 높았습니다
  - 발생 시간: 02시부터 06시까지의 사망률이 가장 높았습니다
  - 연령대: 20대의 교통사고 사고율이 64.4%로 압도적으로 높았습니다
  - 해외 국가 비교: OECD 평균 10만명당 사망자수는 5.1명, 우리나라 인구 10만명당 사망자수는 7.3명으로 평균을 웃돌았습니다
- 머신러닝 분석
  - LSTM: 2015년도~2019년도 서울시 교통사고 시계열 데이터를 활용, 미래에 발생될 서울의 교통사고를 예측하는 모델을 LSTM을 활용하여 만들었습니다. 2015~2018년도 데이터를 훈련용으로, 2019년을 테스트용으로 분리하여, 2019년도에 대한 예측을 시행한

결과, 평균 오차율 15.2%의 모델을 만들 수 있었습니다



- Imbalance Classification: 전체 교통사고중 사망사고를 예측하는 classification을 시행하였습니다. 활용한 데이터셋은 20,21년도 전체 교통사고 대략 39만 6천건이였고, 그 중 사망사고 비율은 1.125%였습니다. 저는 인명사고의 심각성을 고려하였을 때, 모델 평가 척도로서 Precision보다 Recall가 우선시되어야 한다고 판단하였고, Recall 값을 향상시키기 위하여 언더샘플링을 비롯한 전처리를 시행하고, XGBoost와 Logistic 회귀를 통해 분류 예측을 시행하였습니다. 그러나 아쉽게도 만족할만한 Precision-Recall 값을 갖는 binarizer의 threshold 값을 찾을 수 없었습니다.
    - 종속변수와 유의미한 상관관계를 갖는 독립변수가 주어진 독립변수들 내엔 존재하지 않아 이러한 결과가 나온 것이 아닌가 추측하였습니다


Precision and Recall Value
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
precision
recall
-0.04
0.06
0.16
0.26
0.36
0.46
0.56
0.66
0.76
0.86
0.96
Threshold value