# 회고록

**작성자: 권나연, 유영훈**

작성 일**: 2022년 11월 23일**

## 프로젝트 미션

- 비재무 데이터 (외부 데이터)를 활용하여 중소기업의 휴·폐업 예측 모형 구축하는 것

## 고민한 부분 및 해결방법

- 고민한 부분: 처음 예측 모델을 만들었을 때 평가지표 중 AUROC 값이 0.3~0.4 사이의 값을 보였다. 그래서 AUROC 값을 높일 수 있는 방법에 대하여 고민했다.
  - 해결방법: 데이터 셋 분리 시 휴·폐업 기업과 액티브 기업의 비율차이를 고려하였다. 먼저 데이터를 휴·폐업 기업과 액티브 기업으로 나누어 8:1:1 로 데이터를 분리한 후 데이터를 다시 합치는 방법을 이용했더니 AUROC 값이 높아졌다.
- 고민한 부분: 외부 데이터가 실제로 학습에서 중요하게 활용되는지 알 수 없었다.
  - 해결방법: 변수의 중요도를 알 수 있는 트리 기반의 모형들을 활용하였다.

## 아쉬운 점

- 외부 데이터 활용 시 제약이 따랐다. 모델에 코로나19의 영향과 관련된 변수로써 '연간 항공수송실적' 변수를 이용하고 싶었는데, 병합할 수 있는 변수가 상태발생일자 변수뿐이라서 이용할 수 없었다. 날짜 변수에 데이터를 병합하면 $y = 1$ ($y$는 반응변수)인 경우를 나타내는 indicator variable이 되어 모델 예측에 유의미한 영향을 미칠 수 없기 때문이다.
- 개방 데이터의 대부분이 학습에 활용하기 어려운 변수였다. 이름과 관련된 변수거나 분포가 너무 극단적이었다.

## 성장한 부분

- Binary classification을 위한 ML 알고리즘을 실제 데이터에 활용할 수 있게 되었다. 특히 이번 학기 수업에서 배운 BART 모델을 실제 데이터에 이용할 수 있는 기회가 되어 기쁘다.
- 예측값의 분포가 극단적인 상황에서 분류문제를 어떻게 학습해야될지 알게되었다.