Test for Machine Learning Engineer

1. 목표

본 과제의 목표는 승객의 설문 조사 데이터로부터 만족 여부(satisfaction)를 판단하는 것입니다. 먼저 승객의 설문 조사 데이터에 대한 EDA를 수행하고, 이를 바탕으로 적절한 분류 모델을 구현합니다.

2. 데이터

첨부된 압축 파일에는 다음의 4개의 파일이 포함되어 있습니다.

train.csv : 학습 데이터 (10,000개)
→ 일부 데이터가 누락되었습니다. ('Online boarding' 외 10개 열)
unlabeled.csv : 레이블링되어있지 않은 데이터 (99,880개)
→ 일부 데이터가 누락되었습니다. ('Satisfaction' 열)
test.csv : 검증 데이터 (20,000개)
→ 누락된 데이터가 없습니다.
README.txt : 데이터셋에 대한 간단한 설명

위 3개의 csv 파일은 모두 승객 설문 조사 데이터입니다. 하지만 안타깝게도 전체 데이터 중 일부는 satisfaction 항목 등 몇몇 데이터가 누락되었습니다. 아래 표는 각 데이터의 예시입니다.

	train.csv	unlabeled.csv	test.csv
id	84455	101480	126147
Gender	Female	Male	Male
Customer Type	disloyal Customer	Loyal Customer	Loyal Customer
Age	23	47	62
Type of Travel	Business travel	Business travel	Business travel
Class	Eco	Business	Business
Flight Distance	290	3486	3038
Inflight wifi service	3	3	3
Departure/Arrival time convenient	4	3	1
Ease of Online booking	3	3	1
Gate location	3	3	1
Food and drink	5	5	2
Online boarding	NaN	4	3
Seat comfort	NaN	5	3
Inflight entertainment	NaN	4	3
On-board service	NaN	4	3
Leg room service	NaN	4	3
Baggage handling	NaN	4	3
Checkin service	NaN	4	4
Inflight service	NaN	4	3
Cleanliness	NaN	4	4
Departure Delay in Minutes	NaN	2	2
Arrival Delay in Minutes	NaN	0	0
satisfaction	neutral or dissatisfied		neutral or dissatisfied

3. 제출

위 데이터를 기반으로 아래 3가지 항목을 수행하시고 수행과정을 PPT 형식으로 정리하여 코드와 함께 제출해 주시기 바랍니다.

(1) EDA (Exploratory Data Analysis)

어떤 요소가 승객 만족도와 높은 연관성을 가질까요? 누락 여부와 상관없이 전체 데이터(train, test, unlabled)에 대해 EDA를 수행합니다.

(2) train.csv 를 활용한 분류 모델 구현

‘Satisfaction’ 항목이 포함된 데이터만으로 승객의 만족 여부를 예측하는 분류 모델을 구현하고, test.csv로 검증합니다.

(3) train.csv와 unlabeled.csv 를 모두 활용한 분류 모델 구현

‘Satisfaction’ 항목이 누락된 데이터까지 포함해 승객의 만족 여부를 예측하는 분류 모델을 구현하고, 마찬가지로 test.csv로 검증합니다. 단 (2)번에서 구현한 모델보다 더 높은 성능을 낼 수 있는 방법을 찾아봅니다.