비즈니스를 위한 데이터마이닝

2023년 가을학기 중간고사

문제를 풀 때 필요한 가정이 있다면, 어떤 가정하에 문제를 풀었는지 기술하시오.

문제 1)

spotify-2023-2.csv는 2023년에 spotify에서 많이 스트리밍된 음악에 대한 데이터이며, 각 칼럼은 다음과 같다. (총 45점)

컬럼	내용
track_name	노래 제목
artist(s)_name	가수 이름
artist_count	가수 명 수
released_year	출시 년도
released_month	출시 월
released_day	출시 일
in_spotify_playlists	노래가 spotify playlist에 포함된 수
streams	스트리밍 횟수
bpm	노래의 BPM
key	노래의 Key
mode	노래의 모드(Major, Minor)
danceability_%	춤추기 적합 정도
valence_%	긍정 정도
energy_%	인식된 에너지 레벨
acousticness_%	어쿠스틱 사운드 정도
instrumentalness_%	연주 부분의 정도
liveness_%	라이브 공연 정도
speechiness_%	말(대화) 포함 정도

spotify-2023-2.csv를 불러들여, spotify라는 데이터프레임을 만들고 다음을 수행하라.

(1) 종속변수는 스트리밍 횟수인 streams 이며, 독립변수는 데이터 분석가가 탐색하여 선정한다.

(2) 학습집합은 행번호 0~899번까지이며, 나머지 데이터는 테스트 데이터이다.

(3) 선형회귀분석, Ridge, Lasso, ElasticNet을 사용하여 모형을 생성하고, 테스트 데이터에 대한 예측 성과를 측정하라. (성과지표는 RMSE를 사용하라.)

(4) 하이퍼 파라미터 튜닝 및 독립변수 선정과 같은 다양한 방안을 스스로 수행하고, 가장 성과가 좋은 모형을 찾아 모형의 학습방안과 성과를 제시하라.

(성적 채점에 모형을 만드는 방안의 적합성, 변수선정의 적합성과 예측 성과가 반영됨.)

In [ ]:

문제 2)

airline.csv 는 항공사 만족도 데이터이며, 각 칼럼은 다음과 같습니다. (총 50점)

컬럼	내용 (범주는 영어로 표기되어 있습니다.)
Unnamed: 0	행번호
id	탑승구분번호
Gender	성별 Gender of the passengers (Female, Male)
Customer Type	충성 고객/ 비충성 고객 The customer type (Loyal customer, disloyal customer)
Age	나이 The actual age of the passengers
Type of Travel	여행 목적 (개인/비즈니스) Purpose of the flight of the passengers (Personal Travel, Business Travel)
Class	탑승 클래스 (비즈니스, 이코노미, 이코노미 플러스) Travel class in the plane of the passengers (Business, Eco, Eco Plus)
Flight distance	탑승거리 The flight distance of this journey
Inflight wifi service	기내 와이파이 사용 만족도 Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
Departure/Arrival time convenient	출발/도착 시간 편리성 만족도 Satisfaction level of Departure/Arrival time convenient
Ease of Online booking	온라인 예약 만족도 Satisfaction level of online booking
Gate location	게이트 위치 만족도 Satisfaction level of Gate location
Food and drink	음식/음료 만족도 Satisfaction level of Food and drink
Online boarding	온라인 보딩 만족도 Satisfaction level of online boarding
Seat comfort	좌석 편안함 만족도 Satisfaction level of Seat comfort
Inflight entertainment	기내 엔터테이멘트 만족도 Satisfaction level of inflight entertainment
On-board service	온보드 서비스 만족도 Satisfaction level of On-board service
Leg room service	레그룸 서비스 만족도 Satisfaction level of Leg room service
Baggage handling	수화물 처리 만족도 Satisfaction level of baggage handling
Check-in service	체크인 서비스 만족도 Satisfaction level of Check-in service
Inflight service	기내 서비스 만족도 Satisfaction level of inflight service
Cleanliness	청결 만족도 Satisfaction level of Cleanliness
Departure Delay in Minutes	출발 지연 (분) Minutes delayed when departure
Arrival Delay in Minutes	도착 지연 (분) Minutes delayed when Arrival
Satisfaction	항공사 만족도 (만족, 불만족 또는 Neutral) Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

목표변수는 Satisfaction 이고, 입력변수 후보는 나머지 변수들이다.

(1) 입력변수 후보들 중에 불필요한 변수는 제거하라.

(2) 의사결정나무 모형 학습에 적합한 형태로 입력변수들을 변환하여라.

(3) 의사결정나무 모형을 만들고, 5겹 교차검증을 통해 모형의 성과를 측정하고자 한다. 성과측정은 auc 로 측정한다.

(4) 의사결정나무 모형의 max_depth 값을 변화시켜가며, 가장 성과가 좋은 max_depth 값을 찾는다.

(5) (4)에서 찾는 max_depth 값을 활용하여 의사결정나무 모형을 만들고 5겹 교차 검증을 통해 모형의 성과를 측정한다. 성과측정은 auc 로 측정한다.

(6) (5)에서 완성된 모형의 입력변수 중요도를 시각화하라.

(7) (6)에서 파악한 입력변수 중요도를 기준으로 중요도가 높은 10개의 변수를 활용하여, 로지스틱 회귀분석 모형을 만든다. 로지스틱 회귀분석 모형도 5겹 교차검증을 통해 모형의 성과를 측정하고, 성과지표는 auc 를 활용한다.

(8) (5)의 성과와 (7)의 성과를 비교하라.

(성적 채점에 모형을 만드는 방안의 적합성, 변수선정의 적합성과 예측 성과가 반영됨.)

In [ ]:

문제 3) 이 수업에서 개선되어야하는 사항은 무엇인가? (작성 시 5점, 미작성 시 0점)

In [ ]: