비즈니스를 위한 데이터 마이닝

2022년도 가을학기 중간고사

시험 시간: 15시 00분 ~ 16시 40분

- 중간고사 문제 및 데이터는 과제란의 중간고사에 들어 있습니다.
- 다음 페이지부터 있는 문제를 풀어, Jupyter Notebook 또는 다양한 형태의 파일로 중간고사 과제에 제출하세요.
- 문제를 풀 때 필요한 가정이 있다면, 어떤 가정하에 문제를 풀었는지 기술하시오.

HousePrices.csv 는 주택 가격데이터이다. 데이터에 포함된 각 컬럼의 의미는 다음과 같다.

변수	설명
HomeID	주택 번호
Price	주택 가격
SqFt	주택 면적 (단위: 제곱 피트)
Bedrooms	침실 수
Bathrooms	화장실 수
Offers	구매 제안 수
Brick	벽돌 집 여부 (No, Yes)
Neighborhood	동네 (East, North, West)

1. (6점씩 총 42점) 주택 가격(Price)을 예측하는 모형을 아래와 같이 만들려고 한다.

(1) Price를 종속변수로 나머지 변수 중에서 수치형 변수만 독립변수로 하는 회귀분석 모형을 만들어라. 수치형 변수 중 의미가 없다고 생각하는 변수는 제외하여도 되며, 통계치나 회귀분석 모형의 결과를 보고 제외하는 것이 아니라 단순히 변수의 의미만 가지고 판단하여라. 이때, 수치형 변수를 표준화하지 않고 회귀분석 모형을 만들어라. 학습데이터, 테스트데이터로 8:2로 무작위로 구분하고 모형을 통해 테스트데이터에 대한 예측치를 만들고, MAE로 성과를 측정하여라.

(2) (1)과 동일하게 종속변수, 독립변수를 설정하여 회귀분석 모형을 생성하고 10겹 교차검증을 통해 성과를 측정하라. 10겹 교차 검증의 성과지표는 MAE를 활용한다.

(3) Price를 종속변수로 나머지 변수 모두를 독립변수로 하는 회귀분석 모형을 만들어라. 변수 중 의미가 없다고 생각하는 변수는 제외하여도 되며, 통계치나 회귀분석 모형의 결과를 보고 제외하는 것이 아니라 단순히 변수의 의미만 가지고 판단하여라. 이때, 수치형 변수를 표준화하지 않으며 나머지의 경우에는 필요하다고 판단되는 변환 과정을 거쳐라. 위와 같이 회귀분석 모형을 생성하고 10겹 교차검증을 통해 성과를 측정하라. 10겹 교차 검증의 성과지표는 MAE를 활용한다.

(4) (3)에서 만들어진 종속변수와 독립변수를 활용하여, Lasso 회귀분석 모형을 생성하고 성과를 측정하고자 한다. 최적의 alpha 값을 찾기위해 랜덤서치를 활용하며 alpha 값의 범위는 아래의 np.linspace(0.0001, 1000, num=500)을 활용한다. 이는 0.0001부터 1000까지의 값 500개를 균등하게 발생시킨다. 랜덤서치는 200번을 반복하며 성과지표는 neg_mean_absolute_error 를 활용한다.

```
import numpy as np

np.linspace(0.0001,1000,num=500)
```

랜덤서치로 찾은 최적의 alpha 값을 활용하여 Lasso 회귀분석 모형의 성과를 10겹 교차검증으로 측정하라. 성과지표로는 MAE를 활용한다.

(5) (4)에서 수행된 방식대로 Ridge 회귀분석 모형의 최적 alpha 값을 찾고, 이를 활용하여 Ridge 회귀분석 모형의 성과를 10겹 교차검증으로 측정하라. 성과지표로는 MAE를 활용한다.

(6) (3)에서 만들어진 종속변수와 독립변수를 활용하여, Elastic 회귀분석 모형을 생성하고자 한다. (4)와 같은 방식으로 랜덤서치를 수행하며, 다만 alpha는 동일한 방식으로 진행하고 l1 ratio 는 np.linspace(0, 1, num=500) 를 탐색한다. max_iter 값은 100,000으로 설정하여라. 최적 alpha, l1 ratio 값을 활용하여 Elastic 회귀분석 모형을 생성하고 성과를 10겹 교차검증으로 측정하라. 성과지표는 MAE를 활용한다.

(7) (3),(4),(5),(6)에서 얻어진 결과를 이용하여 회귀분석, Lasso, Ridge, Elastic 회귀분석 모형의 성과를 비교하라.

income_evaluation_sub.csv 는 소득수준을 예측하기 위한 데이터이다. income 변수는 소득이 5만 달러 초과인 지(>50K) 그렇지 않은 지(<=50K)에 대한 정보를 담고있다. 각 컬럼은 다음과 같다.

변수	설명
age	나이
educationnum	교육받은 년도
occupation	직종 (Prof-specialty, Craft-repair, ...)
race	인종 (White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other)
sex	성별 (Male, Female)
capitalgain	금융수익
capitalloss	금융손실
hoursperweek	주당 근무시간
income	소득 금액 구분(<=50K, >50K)

2. (5점씩 총 25점) 소득 금액을 구분하는 모형을 다음 단계를 거쳐 나이브 베이스, 서포트 벡터 머신을 이용해 만들고자 한다.

(1) income 컬럼 값이 >50K이면 1, <=50K면 0 값을 갖는 income_category 컬럼을 만들어라.

(2) income_category를 목표변수로, 나머지 변수 중 의미가 없거나 입력변수에 포함되면 안되는 변수를 제외하고 입력변수를 구성하라. 입력변수 중 수치형 변수는 표준화를 수행하고, 범주형 변수는 원 핫 인코딩을 수행하라.

(3) (2)에서 만들어진 데이터를 가지고 나이브 베이스 모형을 만들고자 한다. 나이브 베이스 모형의 성과를 5겹 교차검증으로 측정하라. 이때 성과지표는 f1를 활용한다.

(4) (2)에서 만들어진 데이터를 가지고 서포트 벡터 머신 중 Linear SVC 모형을 만들고자 한다. 하이퍼 파라미터인 C는 하이퍼 파라미터 튜닝을 통해 최적의 모형을 찾아라. 모형의 성과를 5겹 교차검증으로 측정하고, 성과지표는 f1를 활용한다. (LinearSVC는 학습에 시간이 많이 소요되므로, 하이퍼파라미터 튜닝의 횟수를 적게 하여라.)

(5) (2)에서 만들어진 데이터를 가지고 서포트 벡터 머신 중 SVC 모형을 만들고자 한다. 커널 변환은 rbf를 사용하며, gamma는 5, C는 0.001로 설정하여라. 모형의 성과를 5겹 교차검증으로 측정하고, 성과지표는 f1를 활용한다.

BankChurners1.csv 는 한 은행의 이탈 고객과 유지 고객에 대한 데이터이다. 각 컬럼은 다음과 같다.

변수	설명
CLIENTNUM	고객 번호
Attrition_Flag	이탈 여부(Existing Customer/Attrited Customer)
Customer_Age	고객 나이
Gender	성별(M/F)
Dependent_count	가족 수
Education_Level	교육수준 (Graduate/High School/Unknown/Uneducated/College/Post-Graduate/Doctorate)
Marital_Status	결혼상태 (Married/Single/Unknown/Divorced)
Income_Category	소득 수준(Less than $40K, $40K-$60K, $60K-$80K, $80K-$120K, $120K +, Unknown)
Card_Category	카드 등급 (Blue/Silver/Gold/Platinum)
Months_on_book	거래 기간(개월)
Total_Relationship_Count	거래 계좌 수
Months_Inactive_12_mon	지난 12개월동안 거래가 없던 개월 수
Contacts_Count_12_mon	지난 12개월동안 은행 거래 회수
Credit_Limit	신용 대출 한도
Total_Revolving_Bal	총 리볼링 잔고 (대출의 일종이며, 신용카드 사용액을 다음 달에 갚는 것으로 연기)
Avg_Open_To_Buy	지난 12개월동안 대출 잔고
Total_Amt_Chng_Q4_Q1	총 잔액 변화
Total_Trans_Amt	총 거래액
Total_Trans_Ct	총 거래 수
Total_Ct_Chng_Q4_Q1	총 거래 수 변화
Avg_Utilization_Ratio	평균 활용률

3. (총 5점씩 25점) 고객의 은행 이탈을 분류하는 모형을 아래와 같이 만들고자 한다.

(1) Attrition_Flag 변수는 Existing Customer는 0, Attrited Customer는 1 값을 갖는 Flag 변수로 변환하라. Flag는 목표변수이고, Attrition_Flag, Flag, CLIENTNUM를 제외한 나머지 변수 모두가 입력변수이다. 입력변수는 원 핫 인코딩을 수행하며, 수치형 변수의 표준화는 하지 않는다.

(2) (1)에서 만든 데이터를 활용하여 의사결정나무의 최대 나무를 생성하라. 순수도 지표는 gini를 사용하며 10겹 교차 검증을 통해 roc_auc 값의 평균치를 보여라.

(3) (1)에서 만든 데이터를 활용하여 의사결정나무 모형을 생성하라. 이때 max_depth의 값을 2부터 20까지 늘려가면서 모형의 성과를 시각화하라. 순수도 지표는 entropy를 사용하며 10겹 교차 검증을 통해 roc_auc 값의 평균치를 성과로 한다.

(4) (3)에서 가장 좋은 성과를 보인 모형의 입력 변수 중요도를 보여라. 이때, 학습에는 (1)에서 만든 목표변수와 입력변수 전체를 활용하여라.

(5) (4)에서 찾은 입력 변수 중요도에서 중요도 값이 높은 5개의 변수를 입력변수로 하고, (1)에서 만든 목표변수를 목표변수로 하는 로지스틱 회귀분석 모형을 만들어라. max_iter는 10,000을 지정하고 10겹 교차 검증을 통해 roc_auc 값의 평균치를 보여라.

4. 이 수업에서 개선되어야 하는 사항은 무엇인가? (작성 시 8점, 미작성 시 0점)