비즈니스를 위한 데이터 마이닝

2021년도 가을학기 중간고사

시험 시간: 14시 00분 ~ 15시 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.

focus-manual-petrol.csv 는 포드 포커스 (수동, 휘발유) 자동차의 중고가격 데이터이다. Model 변수에는 Focus 하나의 값만 있으며, transmission에도 Manual(수동), fuelType에도 Petrol(휘발유) 하나의 값만 가지고 있다.

변수	설명
model	자동차 모델 (Focus)
year	자동차 생산 년도
price	중고자동차 가격
transmission	자동차 변속기 형태 (Manual: 수동)
mileage	주행거리 (단위: mile)
fuelType	사용 유종 (Petrol: 휘발유)
engineSize	엔진 크기

1. (30점) 포드 포커스 (수동, 휘발유) 자동차의 중고가격(price)을 예측하는 모형을 아래와 같이 만들려고 한다.

(1) price를 목표변수로 나머지 변수 중에서 하나의 값만 갖는 model, transmission, fuelType을 제외한 다른 변수를 입력변수로한다. 입력변수들은 StandardScaler를 통해 표준화하라.

(2) (1)에서 만들어진 입력변수와 목표변수를 활용하여, 회귀분석 모형을 생성하고 5겹 교차검증을 통해 성과를 측정하라. 5겹 교차 검증의 성과지표는 MAE를 활용한다.

(3) (1)에서 만들어진 입력변수와 목표변수를 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 5겹 교차검증으로 측정하라. 성과지표로는 MAE를 활용한다.

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

(5) (1)에서 만들어진 입력변수와 목표변수를 활용하여, Elastic 회귀분석 모형을 생성하고자 한다. (3)과 같은 방식으로 랜덤서치를 수행하며, 다만 alpha는 동일한 방식으로 진행하고 l1 ratio 는 np.linspace(0, 1, num=500) 를 탐색한다.

최적 alpha, l1 ratio 값을 활용하여 Elastic 회귀분석 모형을 생성하고 성과를 5겹 교차검증으로 측정하라. 성과지표는 MAE를 활용한다.

(6) (2),(3),(4),(5)에서 얻어진 결과를 이용하여 회귀분석, Lasso, Ridge, Elastic 회귀분석 모형의 성과를 비교하라.

healthcare-dataset-stroke-data.csv 는 뇌경색 여부와 환자의 건강상태에 관한 데이터이다. 각 컬럼은 다음과 같다.

변수	설명
id	환자번호
gender	성별 (Male/Female)
age	나이
hypertension	고혈압 여부 (0/1)
heart_disease	심장질환 보유 여부 (0/1)
ever_married	결혼 여부 (Yes/No)
work_type	직업 종류 (Private/Self-employed/children/Govt_job/Never_worked)
Residence_type	거주 형태 (Urban/Rural)
avg_glucose_level	평균 혈당 수준
bmi	체질량지수
smoking_status	흡연 상태 (never smoked/Unknow/formerly smoked/smokes)
stroke	뇌경색 발생 유무(1: 발생, 0: 발생하지 않음)

2. (30점) 목표변수는 stroke이며, 뇌경색 발생을 분류하는 모형을 다음 단계를 거쳐 의사결정나무와 랜덤포레스트를 이용해 만들고자 한다.

(1) NA 값이 포함된 행을 제거하라.

(2) 범주형 변수이지만 수치로 표현된 경우에 범주형으로 변환하고, 입력변수와 목표변수로 나누어라. 입력변수에는 원 핫 인코딩을 수행하라.

(3) (2)에서 만들어진 데이터를 무작위로 8:2로 학습집합과 테스트집합으로 나누어라. 학습집합을 가지고 의사결정나무 모형을 생성하고(최대나무, 순수도지표는 entropy), 생성한 모형의 성과를 테스트 집합을 활용하여 측정하라. 성과지표는 정확도를 사용한다.
생성한 모형을 활용하여 각 변수별 중요도를 시각화하라.
의사결정나무 모형의 성과를 5겹 교차검증으로 측정하라. 이때 성과지표는 AUC를 활용한다.

(4) (3)에서 시각화된 의사결정나무 모형의 변수 중요도에서, 가장 중요한 3개 변수를 활용하여 의사결정나무 모형을 생성하라. 최대나무를 만들고 순수도 지표는 entropy를 활용한다. 생성된 모형의 성과는 5겹 교차검증으로 측정하고, 성과지표는 AUC를 활용한다.

(5) (2)에서 만들어진 데이터를 활용하여 랜덤포레스트 모형을 만들어라. n_estimator는 500, max_leaf_nodes는 16, max_features는 auto, max_samples는 0.5를 활용하며, 데이터 중복은 허용한다. 생성된 랜덤포레스트 모형의 성과를 5겹 교차검증으로 측정하여라. 성과지표는 AUC를 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. (30점) 고객의 은행 이탈을 분류하는 모형을 아래와 같이 만들고자 한다.

(1) CLIENTNUM은 제외하고 Attrition_Flag 변수는 Existing Customer는 0, Attrited Customer는 1 값을 갖는 Flag 변수로 변환하라. Flag의 변수형태는 범주형이어야 한다. Flag는 목표변수이고, Attrition_Flag, Flag를 제외한 나머지 변수 모두가 입력변수이다. 입력변수는 원 핫 인코딩을 수행한다.

(2) (1)에서 만든 데이터를 활용하여 LogisticRegression(max_iter=100000), SVC( ), GaussianNB( )를 이용하여 직접 투표 방식으로 결과를 분류하는 VotingClassifier를 만들고 성과를 측정하라. 성과는 10겹 교차 검증을 활용하며, 정확도 지표를 활용하여 측정한다.

(3) (1)에서 만든 데이터를 활용하여 LogisticRegression(max_iter=100000), SVC(probability=True), GaussianNB( )를 이용하여 간접 투표 방식으로 결과를 분류하는 VotingClassifier 를 만들고 성과를 측정하라. 성과는 10겹 교차 검증을 활용하며, 정확도 지표를 활용하여 측정한다.

(4) (1)에서 만든 데이터를 활용하여 GradientBoostingClassifier( )를 만들고자 한다. 최대 깊이는 2로, 분류모형의 수는 500(n_estimators)로, 학습률(learning_rate)은 0.5로 지정하라. 성과는 10겹 교차 검증을 활용하며, 정확도 지표를 활용하여 측정한다.

4. 이 수업에서 개선되어야 하는 사항은 무엇인가? (작성 시 10점, 미작성 시 0점)