2024. 10. 07.

머신러닝과 산업응용 과제1

1. Regression Models

주어진 데이터는 서울시 공공데이터인 자전거 대여 수를 나타낸 데이터이다. 데이터에는 자전거 대여 수와 함께 시간, 날씨 등의 정보를 포함하고 있다. 데이터의 형태는 아래 그림과 같다. Regression model을 통해 자전거 대여 수를 예측하는 것을 목표로 한다. **(Random_State = 42로 고정할 것)**

hour	hour_bef_t	hour_bef_p	hour_bef_v	hour_bef_h	hour_bef_s	hour_bef_c	hour_bef_p	hour_bef_p	count
20	16.3	1	1.5	89	576	0.027	76	33	49
13	20.1	0	1.4	48	916	0.042	73	40	159
6	13.9	0	0.7	79	1382	0.033	32	19	26
23	8.1	0	2.7	54	946	0.04	75	64	57
18	29.5	0	4.8	7	2000	0.057	27	11	431
2	13.6	0	1.7	80	1073	0.027	34	15	39
3	10.6	0	1.5	58	1548	0.038	62	33	23

1) 과제 코드 파일에 작성법을 참고하여 탐색적 데이터분석(EDA)를 진행하시오.

2) EDA를 바탕으로 데이터 전처리 과정(이상치 대체 및 제거, 결측치 대체 및 제거, 스케일링 등)을 진행하시오.

3) 상관계수, VIF 등 변수선택 기법을 자유롭게 선정하여 최종적으로 사용할 변수를 선택하시오.

4) Statsmodels 패키지의 OLS 함수를 사용하여 자전거 대여 수(count)를 예측하는 다중선형회귀(MLR) 모델을 구축하고 테스트 데이터의 자전거 대여 수(count)를 예측하시오.

5) 아래 5개의 lambda 후보에 대해 자전거 대여 수(count)를 예측하는 회귀 모델 두 개(Ridge, Lasso)를 구축하고 **K=10인 K-Fold Cross Validation의 성능을 평균** 내어 아래 표를 작성하여 **분석 보고서에 첨부**하시오. (성능지표로 MSE를 사용하시오.)

모델 \ alpha(λ)	0.001	0.01	0.1	1	10
Ridge					
Lasso					

6) 자전거 대여 수(count)를 예측하는 Decision Tree Regression 모델을 구축하고 과제 코드 파일에 따라 **K=10인 K-Fold Cross Validation**으로 하이퍼파라미터 튜닝을 진행하고 아래 표를 작성하여 **분석**

보고서에 첨부하시오. (성능지표로 MSE를 사용하시오)

하이퍼파라미터 / 모델	max_depth	max_leaf_nodes	max_features	min_samples_split	min_samples_leaf	MSE
Decision Tree Regressor						

7) Ridge, Lasso, Decision Tree Regression모델에 각각 최적의 하이퍼파라미터를 선정하여 최종 예측모델을 구축하고, 테스트 데이터의 자전거 대여 수(count)를 예측하시오.

8) 4번, 7번에서 구축한 Linear Regression (MLR), Ridge, Lasso, Decision Tree Regression 4가지 모델의 테스트 데이터의 자전거 대여 수(count)를 예측한 결과를 아래 **양식에 맞춰** CSV 파일로 제출하시오. (파일명: 학번_이름_regression.csv)

MLR	Ridge	Lasso	Decision Tree Regression
41	…	…	…
…	…	…	…
12	…	…	…

2. Classification Models

주어진 데이터는 와인의 품종 여부를 나타낸 데이터이다. 데이터에는 두개의 와인 품종인 white/red 와인과 다양한 와인 특징 데이터를 포함하고 있다. 데이터 형태는 아래 그림과 같다. Classification model을 통해 와인 품종을 분류하는 것을 목표로 한다. **(red = 1, white = 0), (Random_State = 42로 고정할 것)**

quality	fixed acidi	volatile ac	citric acid	residual su	chlorides	free sulfur	total sulfu	density	pH	sulphates	alcohol	type
5	5.6	0.695	0.06	6.8	0.042	9	84	0.99432	3.44	0.44	10.2	white
5	8.8	0.61	0.14	2.4	0.067	10	42	0.9969	3.19	0.59	9.5	red
5	7.9	0.21	0.39	2	0.057	21	138	0.99176	3.05	0.52	10.9	white
6	7	0.21	0.31	6	0.046	29	108	0.9939	3.26	0.5	10.8	white
6	7.8	0.4	0.26	9.5	0.059	32	178	0.9955	3.04	0.43	10.9	white
6	6	0.19	0.37	9.7	0.032	17	50	0.9932	3.08	0.66	12	white
5	6.1	0.22	0.49	1.5	0.051	18	87	0.9928	3.3	0.46	9.6	white
6	7.1	0.38	0.42	11.8	0.041	32	193	0.99624	3.04	0.49	10	white
5	6.8	0.24	0.31	18.3	0.046	40	142	1	3.3	0.41	8.7	white

1) 과제 코드 파일에 작성법을 참고하여 탐색적 1번 문제와 동일하게 데이터분석(EDA) 및 전처리, 변수선택을 진행하시오.

2) 아래 표와 같이 하이퍼파라미터 후보에 대한 와인 품종(type)을 분류하는 분류 모델 세 개(Logistic Regression, KNN, Decision Tree Classifier)를 만들고 **K=10인 K-Fold Cross Validation**으로 구한 성능지표를 평균내어 아래 표를 작성하여 **분석 보고서에 첨부**하시오. (성능지표로 F1 Score을 사용)

모델 \ C	0.01	0.1	1	10
Logistic Regression				

모델 \ n_neighbors	1	3	5	7
KNN				

모델 \ 하이퍼 파라미터	max_depth	max_leaf_nodes	max_features	min_samples_split	min_samples_leaf	F1
Decision Tree Classifier						

3) Logistic Regression, KNN, Decision Tree Classification 모델에서 각각 최적의 하이퍼파라미터를 사용하여 최종 예측모델을 구축하고 테스트 데이터의 와인 품종(type)을 예측(분류)하시오.

4) 3번에서 구축한 두 가지 모델의 테스트 데이터의 와인 품종(type)을 분류해낸 결과를 아래 **양식에 맞춰** CSV 파일로 제출하시오. (파일명: 학번_이름_classification.csv)

Logistic	**KNN**	**Decision Tree Classification**
1	…	…
…	…	…
0	…	…

3. 제출 자료

1. 분석 보고서 Word 파일**(자신의 분석 과정을 근거를 들어 자세히 설명)**
2. 소스 코드 (코딩이 없는 상용 SW 시 0점 처리)
   → 소스 코드 주석란에 전처리를 어떻게 진행했는지 명시하기 바람
3. 각 문제에 대한 예측 결과 (csv 파일로 제출)
   ★ 주의사항: 코드에서 만든 모델을 실행하여 나온 결과와 불일치 시 0점 처리

4. 평가

1. 회귀 문제는 테스트 데이터의 MSE, 분류 문제는 테스트 데이터의 F1 Score를 통해 최종 평가 순위를 정함 (30%)
2. 분석 보고서 (70%)

5. 제출 기한

- 10월 13일 23시59분 (LearnUs에 제출 자료를 업로드 하시오).
- 이 시간 이후 제출시 5분마다 5점씩 감점

  예) 00시 03분 제출 시 5점 감점

  예) 00시 07분 제출 시 10점 감점