|  |
| --- |
| **SSAC**  **파이썬을 활용한 비즈니스 빅데이터 분석가 양성 과정**  **(TEMPLATE) 자가학습 워크시트** |

|  |  |
| --- | --- |
| **이름** | 백승재 |
| **날짜** | 2021-12-03 |
| **온라인 콘텐츠(인강)**  **학습 범위** | * 직장인을 위한 Python : 예측을 위한 머신러닝 라이브러리 Ch 03.01~03.07 |
| **복습 노트 작성 (자유 양식)** | Ch 03.sklearn – 전처리 기본 – 01. Sklearn 전처리 기능에 대하여 소개  - 데이터를 분석에 적합하게 데이터를 가공/변형/처리/클리닝  (Garbage in / Garbage out)  1) NAN -> Imputer  2) Outlier  3) 정규화 (Normalization) : 0~1사이의 분포로 조정 (분포가 다양하면 모델 성능 떨어짐 방지)  4) 표준화 (Standardization) : 평균을 0, 표준 편차를 1로 맞춤  5) 샘플링 (Over/under Sampling)  6) 피처공학 (Feature Engineering)  - feature 생성 / 연산  - 구간생성, 스케일 변형  Ch 03.sklearn – 02\_train\_test\_split  - 피처/라벨 리스트 생성 -> 데이터 분할(train / validation)  - x\_train,x\_valid, y\_train,y\_valie = train\_test\_split(train[feature],train[label],test=0.2,shuffle=True,random\_state=1)  Ch 03.sklearn –03.빈값처리(Imputer)  - 빈값 확인 : desribe -> isnull().sum()  - 단일 컬럼 빈값 처리 : fillna()  - 여러 컬럼 동시에 처리 : sklearn.impute Import SimpleImputer 사용  - 수치형 데이터와 범주형 데이터와 분리하여 빈값 처리  - 결측값 처리 확인 isnull.sum() -> desribe():  - SimpleImputer(stratege= ‘ ‘) -> fit\_transform , array형태 -> 원래 컬럼에 값 넣어줌  Ch 03.sklearn –04.Label Encoder  - 학습을 위해서 모든 문자로된 데이터는 수치로 변환하여야 합니다.  - 결측 처리 먼저 후에 작업  - 컬럼 1개에 대해서 데이터 수치로 변경 : 함수 정의 -> apply  - sklearn lib 활용 : labelencoder 선언 -> fit\_transrform -> 데이터 프레임에 값 입력  - le.fit\_trasform([train[‘Sex’]]  - 변경 전 값 확인 : le.classes\_  - 변경 전으로 원복 : le.inverse\_transform(train[‘Sex\_num’])  Ch 03.sklearn –05.One Hot Encoding  - 라벨 인코딩으로 처리한 범주형이 이진형이 아닌 경우 머신러닝의 학습 시 각 수치형 값에 대해 연산 관계를 학습해 모델 성능이 떨어지는 경우 발생 -> One Hot Encoding을 이용하여 값을 이진(0,1)로 변환 (컬럼 1개에 대한 변환은 get\_dummies()함수를 사용)  Ch 03.sklearn –06.Nomalization(정규화)  - column간에 다른 min, max값을 가지는 경우(컬럼간에 스케일이 다른 경우), 정규화를 통해 최소치/최대값의 척도를 맞추어 주는 것  - MinMaxSclaer 선언 -> min\_max\_scaler.fit\_transform() -> DataFrame 확인  Ch 03.sklearn –07.Standarization(표준화)  - 데이터가 평균이 0, 표준편차가 1이 되도록 변환  - StandardScaler() 선언 -> fit\_transform |
| **퀴즈 제출** | Q1.다음 회귀 모델의 평가 점수에 대한 설명으로 옳지 않은 것은?  1) MAE는 예측값과 실제값의 차이에 대한 절대값에 대하여 평균을 낸 값이다.  2) MSE는 예측값과 실제값의 차이에 대한 제곱에 대하여 평균을 낸 값  3) RMSE는 MSE에 루트를 씌운 값으로 에러에 따른 손실이 기하 급수적으로 올라가는 상황에서 사용하기 적합하다.  4) R2 score는 RMSE와 동일하게 값이 낮을수록 성능이 좋음을 의미한다.  정답 : 4 : R2Score는 값이 1에 가까울수록 성능이 좋다. |
| **교육생 사진 (교육생 모습 + 교육 화면)** |  |
| **화면 캡쳐**  **(교육 화면 + 교육일시)** |  |