2주차 결과보고서

전공 : 경영학과 학년 : 4학년 학번 : 20190808 이름 : 방지혁

1. 실험시간에 작성한 코드에 전체적인 진행 흐름을 **diagram으로 그리고, 각 단계별로 어떤 기능을 구현했는지 서술하세요**


```
housing = fetch_california_housing(as_frame=True)

data = housing.frame

print(housing.DESCR)
```

1단계는 데이터 세트 불러오기 및 탐색의 과정입니다. Scikit-learn에서 California housing 데이터 셋을 불러와서 data frame 형태로 변환합니다. 데이터에 대한 기본적인 설명도 확인합니다.

```
print(data.shape)

sns.histplot(data['MedHouseVal'])
```

2단계는 데이터 분석 단계로 데이터의 크기를 확인하고 MedHouseVal을 히스토그램으로 시각화해서 데이터의 특성을 파악합니다.

```
X = data.drop('MedHouseVal', axis=1)

y = data['MedHouseVal']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

3단계는 데이터 분할로 독립변수인 X와 종속변수인 y로 분리하고, 전체 데이터를 train_test_split 함수를 활용하여 훈련용 데이터 80%와 테스트용 데이터 20%로 나누어줍니다.

```
model = LinearRegression()

model.fit(X_train, y_train)
```

4단계는 모델 학습 단계로 모델을 생성하고 훈련 데이터를 이용하여 모델이 선형 관계를 학습합니다.

```
train_preds = model.predict(X_train)

test_preds = model.predict(X_test)

train_rmse = np.sqrt(mean_squared_error(y_train, train_preds))

test_rmse = np.sqrt(mean_squared_error(y_test, test_preds))
```

5단계는 모델 평가 단계로 테스트용 데이터에 대해 예측을 수행하고 RMSE를 계산하여 성능을 평가합니다.

2. Linear Regression 모델이 California housing dataset의 변수들을 어떻게 모델링했는지, 수식으로 표현하세요. [HINT] reg.coef_, reg.intercept_

```
for x in model.coef_:

   print(f"{x:4f} ", end = '')

print()

print(f"{model.intercept_:4f}")
```

결과: 0.448675 0.009724 -0.123323 0.783145 -0.000002 -0.003526 -0.419792 -0.433708

-37.023278를 도출할 수 있었습니다.

기존에 print(housing.DESCR)을 했을 때 출력을 보면 :Attribute Information: MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude가 나오는 것을 확인할 수 있습니다. 이를 통해 다음과 같은 수식 MedHouseVal = -37.0233 + 0.4487×MedInc + 0.0097×HouseAge + (-0.1233)×AveRooms + 0.7831×AveBedrms + (-0.000002)×Population + (-0.0035)×AveOccup + (-0.4198)×Latitude + (-0.4337)×Longitude을 도출할 수 있었습니다. 즉 계수가 양수이면 예를 들어 침실이 많을수록 집값이 상승하는 모습을 발견할 수 있었습니다.

3. 결과값으로 나온 **0.7456의 `Test RMSE`값이 적절한 값인지 생각해보고 의견을 서술하세요.**

RMSE는 rooted mean square error로 실제로 측정한 결과 값과 예측 값 간의 차이(오차)를 제곱한 후 평균을 내고 다시 제곱근을 취한 것입니다. 정의에서 알 수 있듯이 해당 값이 작을 수록 예측이 실제와 가깝다는 것을 의미합니다. 스케일을 고려하자면 MedHouseVal의 단위가 100000이기 때문에 실제와 예측치의 차이가 평균적으로 7.4만 달러임을 의미합니다. 또한, Train RMSE인 0.7917과의 차이가 크지 않아서 과적합이 심하지 않다고 생각합니다. 그렇지만 더 줄여 나갈 수 있다고 생각합니다.