Basic Statistics and ML

조재관

1. IRIS Dataset

a. Load dataset and check structure

- Iris 데이터셋을 불러오고 head(), info(), describe(), species 분포를 확인함.

```
species
setosa        50
versicolor    50
virginica     50
```

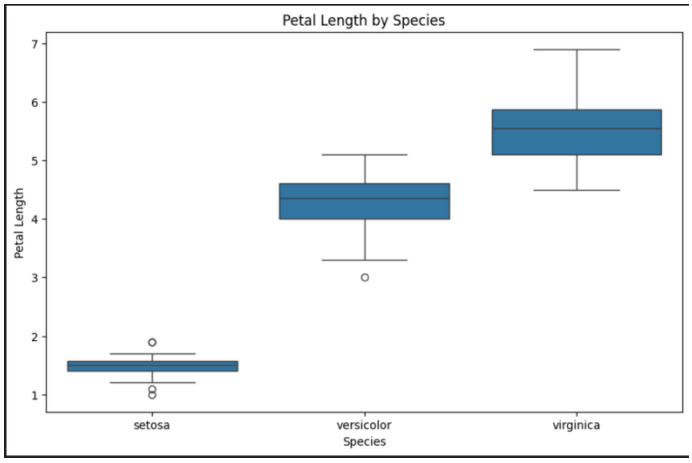
b. Calculating descriptive statics

- Species 별로 Petal Length 에 대해 평균, 표준편차, 사분위수 등을 요약하고 각 클래스별 개수를 확인함.

species	count	mean	std	min	25%	50%	75%	max
setosa	50.0	1.462	0.173664	1.0	1.4	1.50	1.575	1.9
versicolor	50.0	4.260	0.469911	3.0	4.0	4.35	4.600	5.1
virginica	50.0	5.552	0.551895	4.5	5.1	5.55	5.875	6.9

c. Visualization

- Boxplot 을 통해 세 종의 Petal Length 분포 차이를 시각적으로 비교함.



- **Virginica** 는 전체적으로 가장 긴 Petal Length 를 보이고 중앙값도 가장 높음

- **Versicolor** 는 중간 길이의 Petal Length 를 가지고 이상치가 적음

- **Setosa** 는 가장 짧은 Petal Length 를 가지고 몇몇의 이상치들이 보이며, 분포의 범위도 작음.

d. Normality test

- Shapiro-Wilk 검정을 통해 세 종 모두 정규성을 검증

```
setosa     | W-statistic: 0.9550 | p-value: 0.0548 | Normality satisfied
versicolor | W-statistic: 0.9660 | p-value: 0.1585 | Normality satisfied
virginica  | W-statistic: 0.9622 | p-value: 0.1098 | Normality satisfied
```

-세 종률 모두 p-value 가 0.05 이상으로 정규성을 만족함

e. Levene's test

- Levene 검정을 통해 그룹간의 등분산성 검정

귀무가설(H_0): 세 종의 Petal Length 의 분산은 동일하다

대립가설(H_1): 적어도 하나의 종은 분산이 다르다

```
Levene's Test: W-statistic = 19.4803, p-value = 0.0000
```

P-value < 0.05 로 등분산성 가정이 틀림을 확인

f. Hypothesis setting

- ANOVA 를 수행하기 위한 가설 수립

귀무가설(H_0): 세 종의 평균 Petal Length 는 동일하다

대립가설(H_1): 적어도 하나의 종은 평균이 다르다

g. ANOVA Implement

- One-way ANOVA 수행

```
ANOVA: F-statistic = 1180.1612, p-value = 0.0000
```

p-value < 0.05 로 귀무가설을 기각함

h. Tukey's HSD

- Tukey HSD 를 통해 사후 검정을 하여 그룹 쌍 간의 유의미성 확인

```
   Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
  group1     group2   meandiff p-adj lower  upper  reject
-----------------------------------------------------
    setosa versicolor    2.798   0.0 2.5942 3.0018   True
    setosa  virginica     4.09   0.0 3.8862 4.2938   True
versicolor  virginica    1.292   0.0 1.0882 1.4958   True
-----------------------------------------------------
```

세 그룹들 간의 검정 결과 모든 그룹 쌍에서 p < 0.05 로 모두 유의미하다는 것을 확이함

i. Summarizing result

- 이번 분석을 통해 IRIS 의 세가지 종류에서의 Petal Length 평균이 유의미하게 다른지 통계적 방법을 통해서 검증을 하였다. 분석 결과 IRIS 의 Petal Length 의 종에 따라 유의미한 차이를 보이는 것으로 나타났다. Setosa < Versicolor < Virginica 순으로 평균 Petal Length 가 유의미하게 증가한다.

2. Fraud Detection

a. Data load and overview

- 데이터셋을 불러오고, head(), info(), describe()를 통해 구조를 확인한 후 Class 분포를 분석함

```
[Class distribution]
Class
0    284315
1       492
Name: count, dtype: int64

[Class distribution percentage]
Class
0    99.827251
1     0.172749
Name: count, dtype: float64
```

b. Sampling

-사기 거래는 유지하고, 정상 거래는 10,00 건만 무작위로 샘플링하여 새로운 dataset 만듦

```
[Sampled Class distribution]
Class
0    10000
1      492
```

c. Data Preprocessing

-Amount 변수를 StandardScaler 로 정규화하여 Amount_Scaled 로 대체하고, 원래 Amount 변수는 제거

d. Train-Test split

-Stratified split 을 통해 8:2 비율로 데이터를 나누고 클래스 분포를 유지함

```
 Train set distribution
Class
0    7999
1     394
Name: count, dtype: int64

 Train set percentage
Class
0    95.305612
1     4.694388
Name: count, dtype: float64

 Test set distribution
Class
0    2001
1      98
Name: count, dtype: int64

 Test set percentage
Class
0    95.33111
1     4.66889
Name: count, dtype: float64
```

e. SMOTE

-학습 데이터에서 소수 클래스(사기 거래)를 oversampling 하기 위해 SMOTE 를 적용.

SMOTE 적용 이유: 현재의 데이터 셋은 너무 불균형(사기 거래 샘플의 비율이 너무 적음(5% 이하)). 이는 정상 거래의 경우에 대한 Bias 를 생기게 할 수 있고, 사기 거래에서의 패턴을 찾기 어렵게 함. SMOTE 는 인접한 소수 클래스 데이터를 기반으로 synthetic sample 을 생성하여 불균형을 완화함.

```
 Prev distribution of training set
Class
0    7999
1     394
Name: count, dtype: int64

 After SMOTE distribution of training set
Class
0    7999
1    7999
Name: count, dtype: int64
```

f. Model training

- RandomForestClassifier 를 SMOTE 학습 데이터에 학습시키고, 테스트 데이터에 대해 예측을 수행하고 Precision, Recall, F1-score 확인, PR-AUC 계산.

```
              precision    recall  f1-score   support

           0     0.9916    0.9990    0.9953      2001
           1     0.9759    0.8265    0.8950        98

    accuracy                         0.9909      2099
   macro avg     0.9837    0.9128    0.9451      2099
weighted avg     0.9908    0.9909    0.9906      2099


PR-AUC: 0.9158
```

g. Final evaluation

-두 클래스(정상 거래, 사기 거래) 모두에서 Recall ≥ 0.80, F1 ≥ 0.88, PR-AUC ≥ 0.90 를 달성하여 유의미한 모델 성능을 냄. SMOTE 를 적용한 사기 거래의 수치들이 약간 더 낮게 나옴.