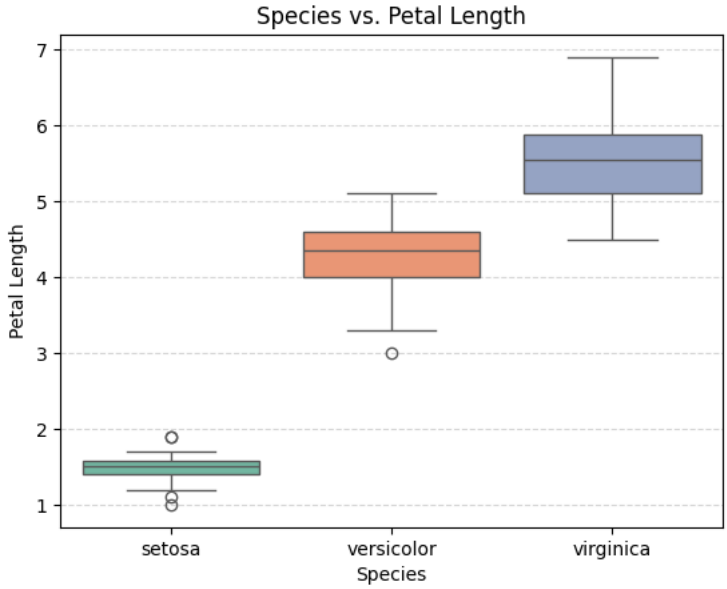
basic_statistics_ML 레포트

27기 이주원

[문제1. Iris 데이터셋을 활용해 클래스별 변수 차이를 검정]

Iris 데이터셋을 불러온 결과, 데이터는 총 150행으로 구성되어 있다. 변수는 'sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'species'이다. 'setosa', 'versicolor', 'virginica' 종이 각각 50개씩 존재함을 알 수 있었다. 또한, 기술통계량을 산출한 결과 'virginica'의 petal length가 5.552로 가장 길며, 'setosa'가 1.462로 가장 짧음을 알 수 있었다.

각 품종의 petal_length 분포를 boxplot으로 시각화하였다.



Boxplot 시각화 결과 virginica의 평균 petal_length가 가장 크다. 또한, setosa의 petal_length의 분산이 제일 작아 분포가 안정적임을 알 수 있다. virginica의 경우 이상치의 범위가 가장 넓음을 확인할 수 있다.

scipy.stats.shapiro()를 활용해 species별 petal_length에 대한 정규성 검정을 수행했다. Shapiro-Wilk 검정의 귀무가설은 '표본 데이터는 정규분포를 따른다.'로, 대립가설은 '표본 데이터는 정규분포를 따르지 않는다.'로 세웠다. setosa의 경우 p-value가 0.0548로, 0.05보다 크므로 귀무가설을 기각할 수 없어 정규성을 만족한다. versicolor의 p-value는 0.1585로, 0.05보다 커서 정규성을 만족한다. virginica는 p-value가 0.1098으로 0.05보다 이 값이 커 정규성을 만족한다.

scipy.stats.levene()를 이용해 세 species 간의 petal_length의 등분산성을 검정했다.

Levene 검정의 귀무가설은 '세 Species는 동일한 분산을 가진다.'로, 대립가설은 '세 Species 중 적어도 한 Species의 분산은 다르다.'로 설정했다. 검정 결과 p-value가 3.1287566394085344e-08로 매우 작게 나와 귀무가설을 기각하여 세 그룹 중 적어도 한 그룹의 분산은 다르다고 결론을 내릴 수 있다.

One-way ANOVA를 실행하기 위해 귀무가설을 '3개 Species 간의 평균의 차이가 유의하지 않다.', 대립가설을 '적어도 하나의 Species의 평균은 나머지와 유의한 차이가 존재한다.'로 수립했다. One-way ANOVA 실행 결과 F통계량이 1180.161이고, p-value가 2.8567766109615584e-91로 0.05보다 작아 귀무가설을 기각하여, 적어도 하나의 Species의 평균은 나머지와 유의한 차이가 존재한다는 결론을 내릴 수 있었다.

One-way ANOVA 실시 결과 세 품종 간 평균에 유의한 차이가 존재하므로 어떤 품종 간 차이가 존재하는지 확인하기 위해 Tukey HSD 사후검정을 실시했다. 그 결과 모든 품종의 p-value가 0.05보다 작아, 모든 쌍 간에 통계적으로 유의한 평균 차이가 존재함을 알 수 있었다.

위에서의 분석을 통해 다음과 같은 결론을 도출할 수 있었다. 첫 번째, Iris 데이터셋의 세 품종 간 petal_length는 모두 정규성을 대체로 만족하고, 등분산성 또한 갖춘 것으로 판단된다. 두 번째, One-way ANOVA 결과 세 품종 간의 평균 차이가 유의미하였으며, Tukey HSD 사후검정을 통해 모든 품종 간 평균에 유의한 차이가 있음을 확인했다. 세 번째, 통계 분석 결과 virginica 품종의 petal_length가 통계적으로 유의하게 가장 길고, setosa가 가장 짧은 것으로 확인되었다.

[문제2. 실제 신용카드 사기 데이터셋을 활용해 클래스 불균형 상황에서 분류 모델을 학습]

신용카드 사기 데이터셋 creditcard.csv는 총 284807행의 데이터로 구성되어 있다. 이 중 정상거래 건수는 284315건으로 약 99.83%를 차지하고, 사기거래 건수는 492건으로 약 0.17%를 차지한다. 즉, 이 데이터셋의 클래스는 불균형하게 구성되어 있으므로 샘플링 및 전처리 과정을 통해 이를 보정해줄 필요성이 있었다.

사기거래(Class=1) 건수는 전부 유지하고, 정상거래(Class=0) 건수는 10000건을 무작위로 샘플링했다. 기존 사기거래 데이터와 샘플링된 정상거래 데이터를 합쳐 새로운 분석용 데이터프레임 sampled_data로 구성했다. 이 데이터셋에서 Class 비율을 다시 확인한 결과, 정상거래 건수는 약 95.31%, 사기거래 건수는 약 4.69%였다.

데이터 전처리 과정으로 Amount 변수를 StandardScaler를 활용해 표준화했다. 이를 Amount_Scaled 변수로 저장했고, Amount 원본 변수를 제거했다. 이후, 종속변수와 독립

변수를 분리하기 위해, y를 sampled_data['Class']로, X를 sampled_data에서 'Class' 변수를 제외한 나머지 변수로 구성했다.

이후, 학습 데이터와 테스트 데이터를 분할했다. train_test_split을 사용해 학습 데이터셋과 테스트 데이터셋을 8:2의 비율로 분할했고, stratify=y 옵션으로 클래스의 비율을 유지하도록 했다. 학습 데이터셋의 경우 정상거래 건수 비율이 약 95.31%, 사기거래 건수 비율이 약 4.69%로 출력됐고, 테스트 데이터셋의 경우 정상거래 건수 비율이 약 95.33%, 사기거래 건수 비율이 약 4.67%로 출력됐다.

원본 데이터셋은 클래스 불균형 문제를 가지고 있어, 분류 모델이 다수 클래스인 정상거래에 편향되어 학습될 가능성이 크고, 이로 인해 소수 클래스인 사기거래를 정확히 예측하지 못할 수 있다. 이러한 문제를 완화하기 위해 SMOTE 기법을 적용하였다. SMOTE는 소수 클래스의 샘플을 기반으로 데이터를 증강해 학습 데이터를 균형있게 만드는 기법이다. SMOTE 기법을 적용해 기존 학습 데이터가 394개였던 사기거래 데이터를 정상거래 데이터와 같은 7999개로 오버샘플링했다.

분류 모델을 구성하기 위한 기본 모델로는 RandomForestClassifier를 선택했다. RandomForest의 경우는 여러 결정 트리를 결합하여 과적합을 방지하고, 예측 성능을 안정적으로 확보할 수 있기 때문이다. 모델을 학습시키고 예측값(predict)와 예측 확률(predict_proba)를 출력했다.

```
Classification Report:
              precision    recall  f1-score   support

           0     0.9945    0.9975    0.9960      2001
           1     0.9457    0.8878    0.9158        98

    accuracy                         0.9924      2099
   macro avg     0.9701    0.9426    0.9559      2099
weighted avg     0.9922    0.9924    0.9923      2099


PR-AUC : 0.9538
```

최종 성능 평가로 classificiation_report, PR-AUC를 분석한 결과, 최종 모델이 Class 0, 1 모두에서 목표치 Recall ≥ 0.80, F1 ≥ 0.88, PR-AUC ≥ 0.90를 달성했음을 확인할 수 있었다. 이는 모델이 불균형한 데이터에서도 클래스를 효과적으로 분류할 수 있음을 의미한다.