데이터마이닝 과제5 – 2022920024 박동찬

1. 실험

Titanic dataset을 Random Forest Classifier를 통해 분류했다. Categorical features(sex, class, deck, embark_town, alone)에 대해 Label Encoding과 One-Hot Encoding을 각각 적용해보고 GridSearchCV를 통해 다양한 hyper-parameter를 탐색했다. 아래와 같은 hyper-parameter 조합을 탐색했다. 또한 5-fold CV를 통해 Train set에서 가장 높은 정확도를 보인 조합을 선택하였다.

```
param_grid = {
    'n_estimators': [25, 50, 100, 125, 150, 200],
    'criterion': ['gini', 'entropy'],
    'max_depth': [None, 2, 4, 6, 8, 10, 15, 20],
    'min_samples_split': [2, 3, 4, 5, 6, 8, 10, 15, 20],
    'min_samples_leaf': [2, 4, 5, 6, 8, 10],
    'max_features': ['sqrt', 'log2', None]
}
```

2. 실험 결과

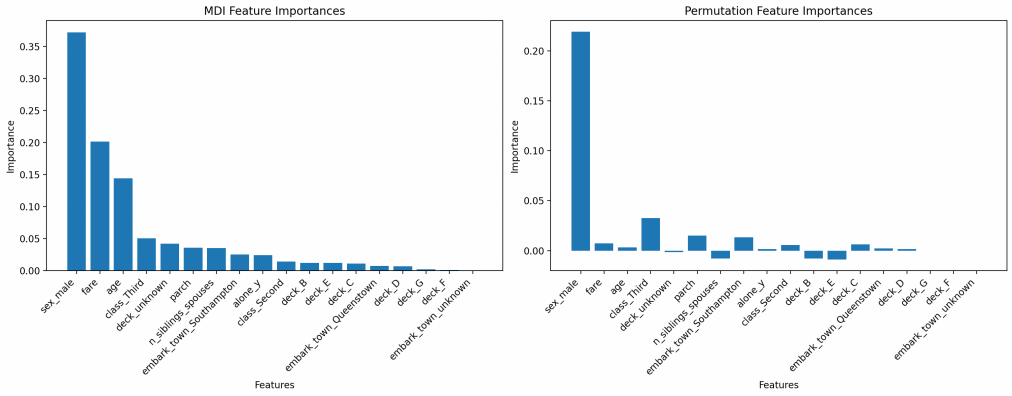
A. Best Accuracy Setting

```
----- Encoding Method: label -----
Fitting 5 folds for each of 15552 candidates, totalling 77760 fits
Best params: {'criterion': 'gini', 'max_depth': 10, 'max_features': None, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 125}
CV Best Accuracy: 0.8304
Test Accuracy: 0.8016

----- Encoding Method: onehot -----
Fitting 5 folds for each of 15552 candidates, totalling 77760 fits
Best params: {'criterion': 'gini', 'max_depth': 15, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 50}
CV Best Accuracy: 0.8303
Test Accuracy: 0.8095

========================================
Best performing encoding method: onehot
Test Accuracy: 0.8095
Execution time: 19844.481937s
```

B. Feature importance(MDI, Permutation)



3. Discussion

A. Encoding Methods

Label Encoding으로 학습한 Random Forest의 Test Accuracy는 0.8016, One-Hot Encoding으로 학습한 Random Forest의 Test Accuracy는 0.8095가 도출되었다. Titanic dataset에는 Categorical features(sex, class, deck, embark_town, alone)가 존재하는데 특정 feature의 경우, 순서 정보가 없음에도 불구하고 Label Encoding 방식은 정수 값을 부여하면서 암묵적인 순서

정보가 들어갈 수 있다. 이로 인해 오류가 나타날 수 있고 One-Hot Encoding은 순서 관계 없이 모델에게 정확한 정보를 학습시킬 수 있어 더 좋은 일반화 성능을 보인 것이라고 할 수 있다.

B. N_estimators

N_estimators는 Random Forest를 구성하는 개별 의사결정트리의 개수를 의미한다. 트리가 많아질수록 일반적으로 모델의 안정성과 성능이 올라가고 훈련 시간이 증가한다. 본 실험에서는 Label Encoding에서 125, One-Hot Encoding에서 50이 나왔는데 인코딩 방식과 데이터 분할 상황에서 모델이 Overfitting 되는 경향이나 트리 개수가 늘어날 때 정확도가 크게 향상되지 않는 상황 등이 복합적으로 영향을 준 것으로 볼 수 있다.

C. Criterion Methods

두 가지 인코딩 방식 모두 gini impurity method가 선택되었다.

D. Max depth

두 인코딩 방식 결과에서 각각 10, 15로 도출되었는데 이는 트리의 깊이가 너무 얕으면 underfitting이 발생하고 너무 깊어지면 overfitting이 발생할 수 있다. 따라서 중간 정도의 적절한 깊이 제한이 모델의 일반화 성능을 높인 것으로 볼 수 있다. 또한 One-Hot Encoding 방식은 범주형 변수의 차원이 증가하므로 트리가 상대적으로 복잡해질 필요가 있어 Label Encoding에 비해 더 깊은 트리에서 최고 성능을 보인다.

E. Min samples split

두 인코딩 방식 결과에서 각각 5(label encoding), 2(one-hot encoding)로 도출되었다. 값이 작을수록 노드가 쉽게 분할되어 깊은 트리를 만들고 값이 클수록 트리가 얕아지는 경향을 보인다. Label Encoding 방식은 비교적 중간 정도의 값에서 최적의 성능을 보였고 One-Hot Encoding의 경우에서는 매우 빠르게 노드를 분할해서 세분화된 규칙을 찾았는데 앞서 언급한 max_depth=15와 결합되어 범주가 많아진 데이터를 충분히 분할하는 방향으로 학습되었다고 볼 수 있다.

F. Min samples leaf

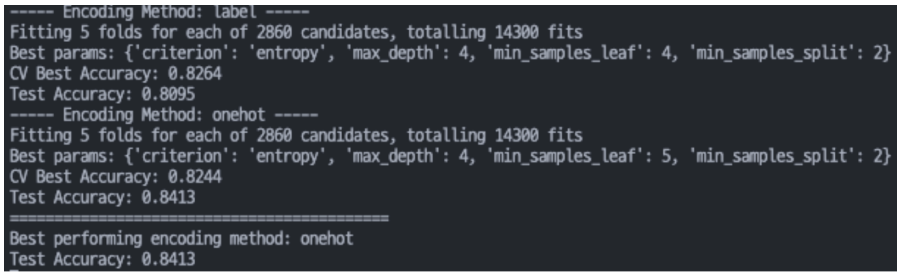
두 인코딩 방식 결과에서 각각 2(label encoding), 2(one-hot encoding)로 도출되었다. 분할 후 leaf node에 존재해야 하는 최소 샘플 수로써 너무 작은 값을 사용하면 leaf node에 작은 샘플만 포함해서 noise에 민감해지고 너무 큰 값을 사용하면 세밀한 부분을 학습하지 못하여 모델의 표현력이 감소한다. 이 경우에서는 비교적 작은 값에서 최적의 성능을 보여 overfitting의 가능성을 생각해볼 수 있다.

G. Max_features

Max_features는 각 분할 시 사용할 특성 개수의 최대치를 의미한다. 두 인코딩 방식 결과에서는 각각 None(label encoding), sqrt(one-hot encoding)가 최적의 성능을 보였다. 여기서 None은 모든 feature를 사용하는데 Label Encoding에서는 모든 feature를 고려해도 크게 overfitting 되지 않았음을 의미할 수 있고 One-Hot Encoding은 차원이 많아져 sqrt로 제한하는 것이 더 효과적이었을 것이라 생각할 수 있다.

H. Decision Tree와의 성능 비교

Decision Tree Classifier를 사용한 결과는 다음과 같다.



일반적으로 Random Forest가 단일 Decision Tree보다 성능이 높거나 유사한 경우가 많지만 실제 실험 결과에서는 단일 Decision Tree의 Test Accuracy가 더 높게 도출되었다. 이는 데이터 분할의 무작위성, 파라미터 검색 범위 및 초기 조건 등에 따라 발생할 수 있는 일종의 예외 상황이라고 볼 수 있다. 이번 실험에서 매우 많은 파라미터 조합(155520개)을 탐색하긴 했으나 특정 구간에서 다른 최적점을 놓쳤거나 현재의 데이터 분할(seed)에 따라 단일 트리가 더 좋은 일반화 성능을 낸 것일 수도 있다.

I. Feature importance(MDI vs Permutation)

i. MDI(Mean Decrease Impurity)

Random Forest가 내부적으로 트리 분할에 기여한 불순도 감소량을 합산해 중요한 feature를 파악한다. 빠르게 확인할 수 있지만 여러 feature가 상관관계를 가질 때 중요도가 왜곡될 수 있다. 위 결과 그래프에서 sex_male이 가장 생존율에 중요한 영향을 미치는 것을 도출되었다. 그 다음으로 age, fare, class_Third, deck_unknown이 높았다.

ii. Permutation Importance

학습된 모델에 대해 특정 feature 값을 무작위로 섞어보며 모델 성능이 얼마나 떨어지는지를 기준으로 중요도를 평가한다. 실제 예측 성능에 미치는 영향을 직접 측정하기에 보다 정확한 지표로 볼 수 있지만 계산 비용이 더 크다는 단점이 존재한다. 위 결과 그래프에서 sex_male이 가장 크게 기여하는 모습을 보이고 있다. 그 다음으로 class_Third, parch, embark_town_Southampton이 높았다. MDI와 sex_male이 가장 크게 기여하는 것은 같지만 그 다음 feature는 다르게 나타났다. 이는

요금이 높은 사람일수록 높은 등급의 객실을 사용하고 갑판 정보가 비교적 많은 상관관계가 존재한다. 이렇게 다른 feature와의 상관관계로 인해 MDI가 과도하게 측정되었을 가능성이 존재한다.