의사결정 나무 실습 과제

<에이쁠원하죠>

1

Data & preprocessing

-범주형 종속변수 분석에 사용된 Data: https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets

고객의 데이터와 각 고객의 금융 상품 구매 유/무를 기록한 데이터이다. 나이, 직업, 결혼 여부 등 다양한 독립 변수들이 존재하고 우리는 나이, 결혼 여부, 부동산/일반 부채 소유 여부의 4가지 변수를 선택하여 분석을 진행하였다. 결측치가 Unknow으로 기록되어 있기 때문에 따로 제거를 해주었다.

-연속형 종속변수 분석에 사용된Data:
https://www.kaggle.com/datasets/amineoumous/50-startups-data?select=50_Startups.csv

미국 스타트업 기업의 데이터와 각 기업의 이익을 기록한 데이터이다. R&D 투자, 기업 관리 투자, 마케팅 투자, 소속된 주가 독립 변수로 존재하고 우리는 이 중 소속된 주를 제외하고 분석을 진행하였다.

2

C4.5, CART, CHAID

C4.5는 이진 분기를 사용하여 트리를 생성하며, 분기 기준으로는 Information gain을 사용한다. Information gain은 분기 전과 후의 데이터 집합의 엔트로피를 측정하여 분기 기준을 선택하는 방식이다.

- 엔트로피 $E=-\sum ki=1pilog2(pi)$

CART(Classification and Regression Tree)는 CART는 이진 분기(Binary Split)를 사용하여 트리를 생성하며, 분기 기준은 주로 지니 계수(Gini Index)를 사용한다. 지니 계수는 분기 전과 후의 데이터 집합의 불순도를 측정하여 분기 기준을 선택한다. 목표변수가 범주형인 경우에는 Classification tree, 연속형인 경우에는 Regression tree를 이용한다. 연속형 목표변수에 대한 분리기준으로는 Variance reduction과 F-test(평균 차이 검정)을 사용한다. CART는 과적합 문제를 해결하기 위해 가지치기(Pruning)를 사용한다.

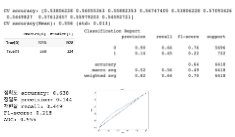
- 지니계수 $Gini(t,D)=1-\sum l\in levels(t)P(t=l)2$

CHAID(Chi-square Automatic Interaction Detector)는 범주형 데이터를 사용하는 분류 문제에 적합한 DT 알고리즘이다. CHAID는 카이제곱 검정(Chi-square Test)을 사용하여 분기 기준을 결정합니다. 카이제곱 검정은 속성 간의 관련성을 측정하여 가장 관련성이 높은 속성을 기준으로 분기한다. CHAID는 가지치기 기법을 사용하여 과적합 문제를 해결한다.

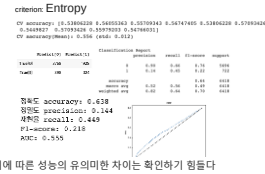
3

Classification DT는 범주형 목표변수에 대한 의사결정나무기법으로, 분류하고자 하는 대상을 분류 기준에 따라 클래스(class)라는 카테고리로 나누어 분류하여 트리 형태의 구조를 만든다. 트리 구조의 각 노드(node)에서는 해당 속성의 값을 비교하여 다음 노드로 분기하게 된다.

criterion: Gini index



criterion: Entropy



-> 정확도, 정밀도, 재현율, ROC커브 등에 대하여 criterion 차이에 따른 성능의 유의미한 차이는 확인하기 힘들다

- 정확도(Accuracy): 전체 예측 결과 중에서 올바르게 예측한 결과의 비율을 나타내는 지표. (TP+TN)/(TP+FP+FN+TN)으로 계산된다.
- 정밀도(Precision): 모델이 Positive로 예측한 결과 중에서 실제로 Positive인 결과의 비율을 나타내는 지표. TP/(TP+FP)으로 계산된다.
- 재현율(Recall): 실제 Positive인 샘플 중에서 모델이 Positive로 예측한 결과의 비율을 나타내는 지표. TP/(TP+FN)으로 계산된다
- F1 점수(F1 Score): 정밀도와 재현율의 조화 평균을 나타내는 지표. 2*((precision*recall)/(precision+recall))으로 계산된다.
- AUC(Area Under the Curve)는 ROC(Receiver Operating Characteristic) 커브 아래 면적을 나타내는 지표. 1에 가까울수록 좋은 성능을 나타내며 AUC가 0.5에 가까우면 모델의 성능이 무작위 수준에 가깝다는 것을 의미한다.
- ROC(Receiver Operating Characteristic) 커브: x축과 y축에 FPR(False Positive Rate)과 TPR(True Positive Rate)을 나타

4

Optimization

max_depth

max_leaf_nodes

-> 1. max_leaf_nodes를 비교한 경우에는 traing accuracy와 validation accracy의 차이가 max_depth를 비교한 경우에 비해 미미하므로 max_depth를 `hyper parameter`로 설정

2. Grid Search 결과
grid search에 사용된 hyper-parameter는 위의 max_leaf_nodes와 같이 큰 의미를 보이지 않는 파라미터를 제외하고 선택하여 튜닝했다.

```
gs.best_score_:  0.5789194197405744

gs.best_params_:
{'decisiontreeclassifier__criterion': 'entropy',
 'decisiontreeclassifier__max_depth': 6,
 'decisiontreeclassifier__min_samples_leaf': 30,
 'decisiontreeclassifier__min_samples_split': 10}
```

3. 성능 확인

	Predict[0]	Predict[1]
True[0]	4559	1137
True[1]	478	244

```
정확도 accuracy: 0.745
정밀도 precision: 0.1764
재현율 recall: 0.3380
f1-score: 0.2318
```

AUC: 0.5627
Optimization을 하기 전보다 정확도는 0.1 이상 증가했고, 정밀도는 약 0.03 증가하였으며, F1 점수는 0.012 증가하였고, AUC는 0.022 증가하여 전반적 평가지표를 통해 성능이 증가하였음을 확인할 수 있다.

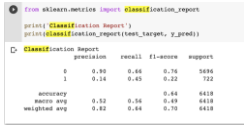
<최종 DT>
종이로운 규칙은 연령대에 따라 60세 이상 고령 고객과 결혼을 하지 않은 20대 초반의 금융상품 구매율이 높았다는 것이다. 이 결과를 이용하면 다른 금융상품의 마케팅 성공률을 높힐 수 있을 것이라 생각된다.
아래 Feature importance를 참고해봐도 연령대가 가장 중요한 변수로 작용하고 있는 것을 확인할 수 있다.

5

Regression & Classification DT

Regression DT는 데이터를 분석하여 입력 변수와 출력 변수 간의 관계를 모델링하는 의사결정나무의 종류 중 하나
의사결정나무는 특성 공간을 분할하는 분할규칙을 통해 데이터를 분류하는 데 사용되며, 회귀 결정 트리는 예측값을 출력하는 데 사용
특성과 관계를 쉽게 이해할 수 있고 데이터 전처리가 거의 필요하지 않다는 장점이 있으나 과적합 문제가 발생할 수 있음

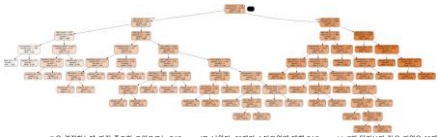
Classification DT
class 0의 precision은 0.9, class 1의 precision은 0.14로 class 0에 대해서는 90%의 샘플이 negative, class 1에 대해서는 14%의 샘플이 positive였음을 의미
class 0의 recall은 0.66, class 1의 recall은 0.45로 class 0에서 66%의 샘플이 negative였지만 모델이 positive로 예측했고, class 1에서는 45%의 샘플만이 positive였고 모델 또한 positive로 예측했음을 의미.
class 0에 대한 f1-score는 0.76, class 1에 대한 f1-score는 0.22
support는 각 class에 대한 실제 샘플 수로 class 1은 5696, class 2는 722개
macro avg의 precision, recall, f1-score의 macro avg가 0.52, 0.56, 0.49인 것을 모았을 때 두 클래스에 대한 분류 성능이 균등하지 않은 것으로 보임. class 1의 예측 성능을 높이면 개선할 수 있을 것
weight avg에 해당하는 값 또한 동일한 양상을 띔. 따라서 class 0의 예측 성능이 class 1에 비해 더 우수하다고 할 수 있음



```
from sklearn.metrics import classification_report

print('Classification Report')
print(classification_report(test_target, y_pred))
```

```
Classification Report
              precision    recall  f1-score   support

           0       0.90      0.66      0.76      5696
           1       0.14      0.45      0.22       722

    accuracy                           0.64      6418
   macro avg       0.52      0.56      0.49      6418
weighted avg       0.82      0.64      0.70      6418
```

6



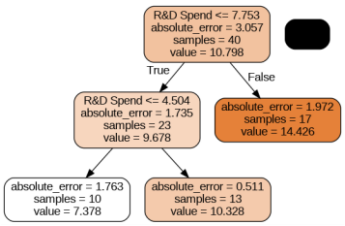
profit을 결정하는데 가장 중요한 요인으로는 R&D spend로 나왔다. 50개의 스타트업에 대해 R&D spend가 7만 달러보다 작은 기업은 25개, 큰 기업은 17개로 대부분의 스타트업이 R&D spend가 7만 달러보다 적었음을 알 수 있다. R&D spend가 7만 달러보다 작은 기업 중 R&D spend가 500 달러보다 적은 기업의 경우 administration이 다음으로 중요한 변수였다. R&D spend가 2만달러 이상,4만달러 이하인 기업의 경우 marketing spend가 그 다음으로 고려하는 변수였다. R&D spend가 4만 달러 이상, 7만 달러 이하인 기업의 경우 sample 수가 다르므 가장 말았다. 이상우 administration과 R&D spend를 순차적으로 고려하는 것을 알 수 있다.
R&D spend가 7만달러 이상인 스타트업의 경우 이때 R&D spend가 14만 달러보다 큰 기업은 한개였고 14만 달러보다 적은 키업의 경우 marketing spend가 그다음으로 고려되었다. R&D spend가 12만달러 이상, 13만 달러 이하인 경우 그 다음으로 administration을 고려했다. 이때 administration이 11만 달러 이상 12만 달러 이하인 경우 administration만이 중요한 고려 대상이였고 11만 달러 이하의 기업의 경우 다음으로 R&D spend를 고려했고 R&D spend가 9만 달러 이상이면 다음으로 marketing spend를 고려했다.
종합적으로 보아 R&D spend가 500만 달러 이상 14만 달러 이하인 경우 다음으로 marketing spend가 중요하게 작용하는 것 같다.

```
R squared: 0.794
MSE: 3.323
```

r squared값과 MSE값을 보아 DT의 성능이 괜찮은것 같다.

7

Best model



Grid Search 결과
grid search에 사용된 hyper-parameter는 max_leaf_nodes와 같이 큰 의미를 보이지 않는 파라미터를 제외하고 선택하여 튜닝했다.

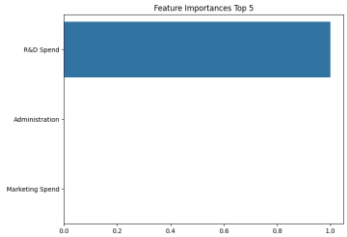
최종 Score와 선택된 hyper-parameter는 다음과 같다.

gs.best_score: 5.380773600946194

'decisiontreeregressor__criterion': 'absolute_error',
'decisiontreeregressor__max_depth': 2,
'decisiontreeregressor__min_samples_leaf': 10,
'decisiontreeregressor__min_samples_split': 10

최적의 파라미터로 모델을 생성한 결과 profit을 결정하는데 중요하게 적용되는 변수는 오직 R&D spend인 것으로 확인되었다.

8



best model을 이용해 중요한 features를 확인한 결과 의사결정나무에서 확인한 것과 마찬가지로 R&D spend가 가장 중요한 것으로 나왔다다

9

한계점 및 개선점

-한계점:

데이터 전처리 과정에서 분석하기 용이하도록 임의로 독립변수를 제거하였다. 그에 따라 모델의 성능 저하가 발생할 수 있다. 실제로 연속형 종속변수를 예측하는 모델의 결과에서 Feature importance가 1개의 변수에 편향되어 나타났다. 인과관계가 명확하지는 않지만 전처리 과정이 그 원인일 수 있다. 혹은 연속형 종속변수에 대한 데이터 개수가 적은 것도 원인일 가능성이 있다.

-개선점:

모든 변수에 대하여 분석을 진행하기 위해서 범주형 독립변수를 Dummy 변수를 이용해서 포함시켜 분석할 수 있다. 또한 연속형 종속변수에 대한 데이터 개수를 늘리는 것도 연구를 개선시킬 수 있는 방법이다. 그리고 Grid search 과정에서 hyper-parameter를 4개정도 선택해서 조합을 분석했는데, 더 중요하거나 많은 종류의 hyper-parameter 조합을 비교하면 모델의 정확도가 향상될 가능성이 있다.

10