2024.08.30 머신러닝 실습 보고서

Created	@2024년 8월 30일 오후 3:29
Class	8.27 ~ 9.3 모델 평가

이름	곽태경
소속	빅데이터 9기

요약

8월 30일 실습에서는

SNS를 이용자 정보를 담은 데이터셋에서 나이와 예상 소득을 특징(feature)으로 두고, 광고 상품 구매 여부를 목표 변수(target)으로 설정했습니다.

그리고 6가지의 머신러닝 모델 중 Stratified k-Fold 교차 검증을 통해 가장 적합한 모델을 선별했습니다.

그 결과는 Random Forest Classifier로 나왔고, Random Forest Classifier를 GridSearchCV를 이용해 최적의 하이퍼파라미터를 찾았습니다.

그리고 찾은 값을 토대로 모델을 학습시켜 나이와 예상 소득에 따른 광고 상품 구매 여부를 예측하는 모델을 구현했습니다.

‘그리고 모델의 정확도와 혼동 행렬을 계산하고, ROC, 학습 곡선으로 모델 검증 과정을 거쳤고, SHAP를 이용해 설명 가능한 인공지능(XAI) 기법으로 feature들의 영향력도 시각화해보았습니다.

그리고 결과를 통해 비즈니스 측면에서의 예측 모델 사용 전략을 제시해 보았습니다.

도입

문제 정의

SNS 사용자들은 서비스를 이용할 때 개인 맞춤 광고를 자주 접하게 됩니다. 그런데, 광고의 과도한 노출과 비효율적인 타겟팅으로 인해 사용자 만족도가 떨어지는 문제가 있는데, 이러

한 문제를 해결하지 않으면 사용자들은 지겨움을 느끼고, 광고주들은 투자 대비 낮은 수익을 얻게 될 위험이 있습니다. 이 문제를 해결하려면, SNS 사용자 정보들로 광고 상품 구매 여부를 예측해 보는 것이 해결 방법이 될 수 있을 것입니다. 만약 성공적으로 예측이 가능하다면, 광고의 양이나 사용자 맞춤 광고의 대상을 조정해볼 수 있을 것입니다.

기획 의도

- 머신 러닝 모델을 사용하여 SNS 사용자들의 나이, 예상 소득 정보로 광고 상품 구매 여부를 예측하고, 이를 통해 광고 효율성을 극대화할 수 있는 전략 제시를 위해서입니다.

개발 환경

- Vscode 1.92.2
- Anaconda 1.12.3
  - Python 3.12.4
  - Numpy 1.26.4
  - Pandas 2.2.2
  - matplotlib 3.8.4
  - Scikit_learn 1.4.2
  - Seaborn 0.13.2
  - jupyter notebook v2024.7.0

데이터 정보

파일명 : Social_Network_Ads.csv

데이터 정보

분류	값(개수)
sample	400
feature	5

feature 목록 (결측치 : 없음)

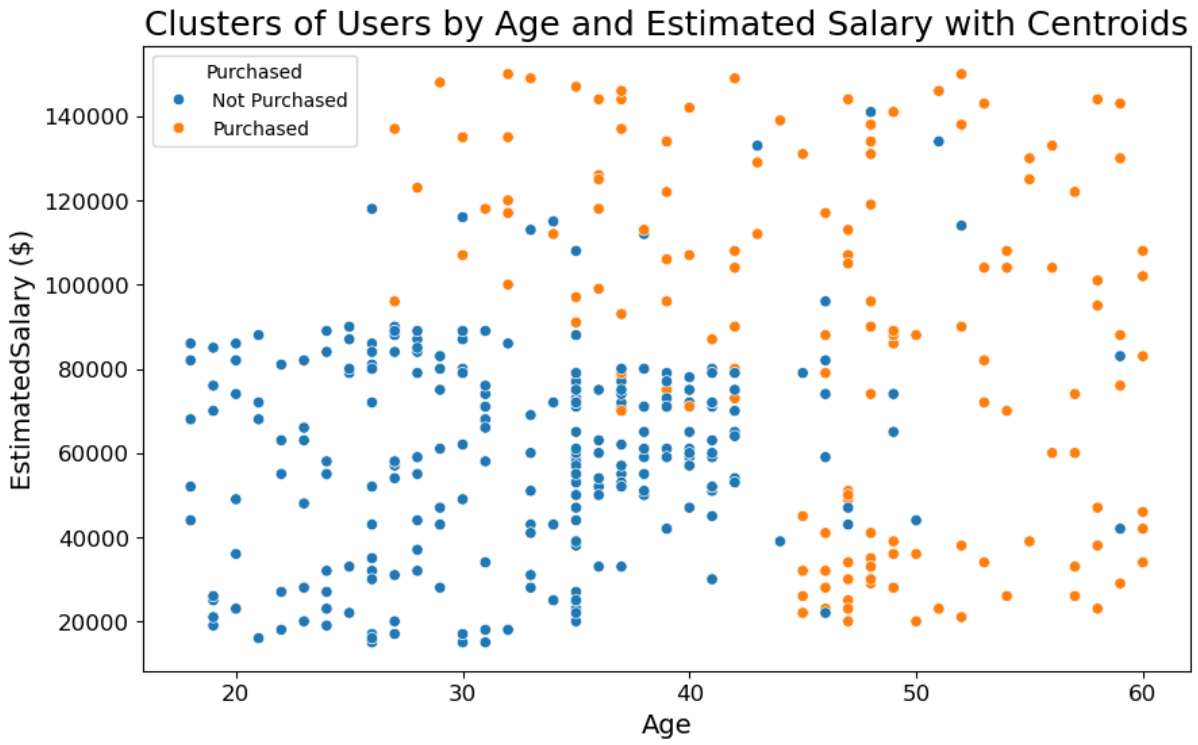
features	정보	데이터 타입	데이터 예
User ID	사용자 id	int64	15624510
Gender	사용자 성별	object	Male
Age	사용자 나이	int64	18
EstimatedSalary	사용자 예상 소득($)	int64	15000
Purchased	광고 상품 구매 여부	int64	0

개발 과정

1. EDA

'User ID' feature는 사용자의 고유 id인데, 400개의 샘플이 모두 다른 사람의 고유 정보이기 때문에 인덱스로 구분해도 된다고 판단하여 열 삭제를 하였습니다.

그리고 이상치를 알아보기 위해서 결제 여부로 구분한 나이와 예상 소득의 관계 산점도를 시각화 해 보았습니다.



이상치 고려

그래프를 통해 잠재적인 이상치를 검토한 결과, 일반적인 상식에 비추어 "25세 이하, 예상 소득 10만 달러 이상"을 이상치로 간주했습니다. 이는 보편적인 사회적 기준에서 벗어나는 것으로 판단했기 때문입니다. 그래프에 따르면, 러한 조건에 해당하는 데이터는 존재하지 않았으므로, 별도의 이상치 제거 과정은 필요하지 않았습니다.

전처리

먼저 feature와 target을 설정했습니다.

features	정보
Age	사용자 나이
EstimatedSalary	사용자 예상 소득

target	정보
Purchased	광고 상품 구매 여부

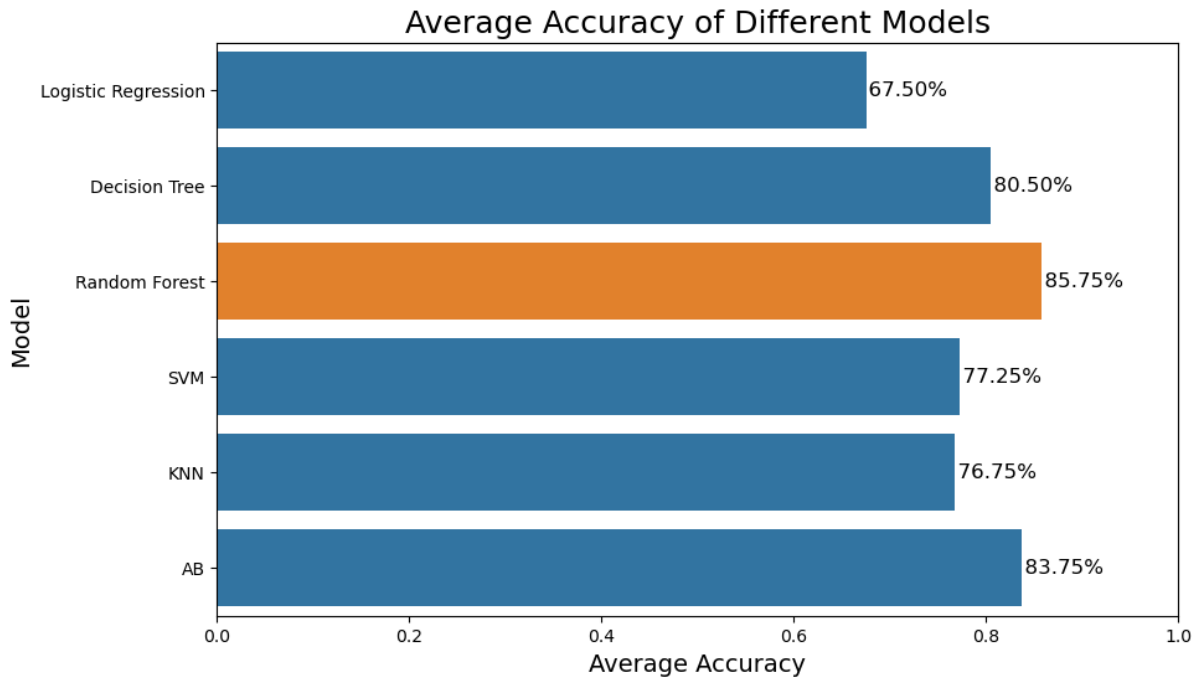
그리고 feature들을 StandardScaler로 표준화하여, 표준 정규 분포를 따르는 데이터로 변환하여 모델 학습의 효율성을 높였습니다.

2. 모델 선별 및 학습

그리고, 6가지 모델을 Stratified K-fold 교차 검증을 5회 반복하여 정확도를 평가해서 정확도의 평균이 가장 높은 모델을 선별했습니다.

(모델 random_state : 12345, K_fold Shuffle = False)

모델명	정확도 평균
Logistic Regression	67.50%
Decision Tree	80.50%
Random Forest Classifier	85.75%
SVC (SVM)	77.25%
K-neighbors classifier(KNN)	76.75%
AdaBoostClassifier(AB)	83.75%



식별 결과, Random Forest Classifier가 정확도가 85.75%로 가장 높았으므로, 이 모델을 선택하여 분류 모델을 구현해보기로 결정했습니다.

모델 교차 검증 과정에서 선택한 **RandomForestClassifier**를 **GridSearchCV**를 사용하여 5회 교차 검증을 실시했습니다. 이 과정에서 모델의 일반화 성능을 최적화하기 위해 교차 검증 데이터의 정확도 평균이 가장 높은 하이퍼파라미터를 탐색했습니다.

최적의 하이퍼파라미터	교차 검증 샘플 정확도 평균
min_samples_leaf: 4, n_estimators: 200	0.918750

그리고, 전체 샘플을 학습용 샘플과 테스트용 샘플로 8:2로 나누었습니다.

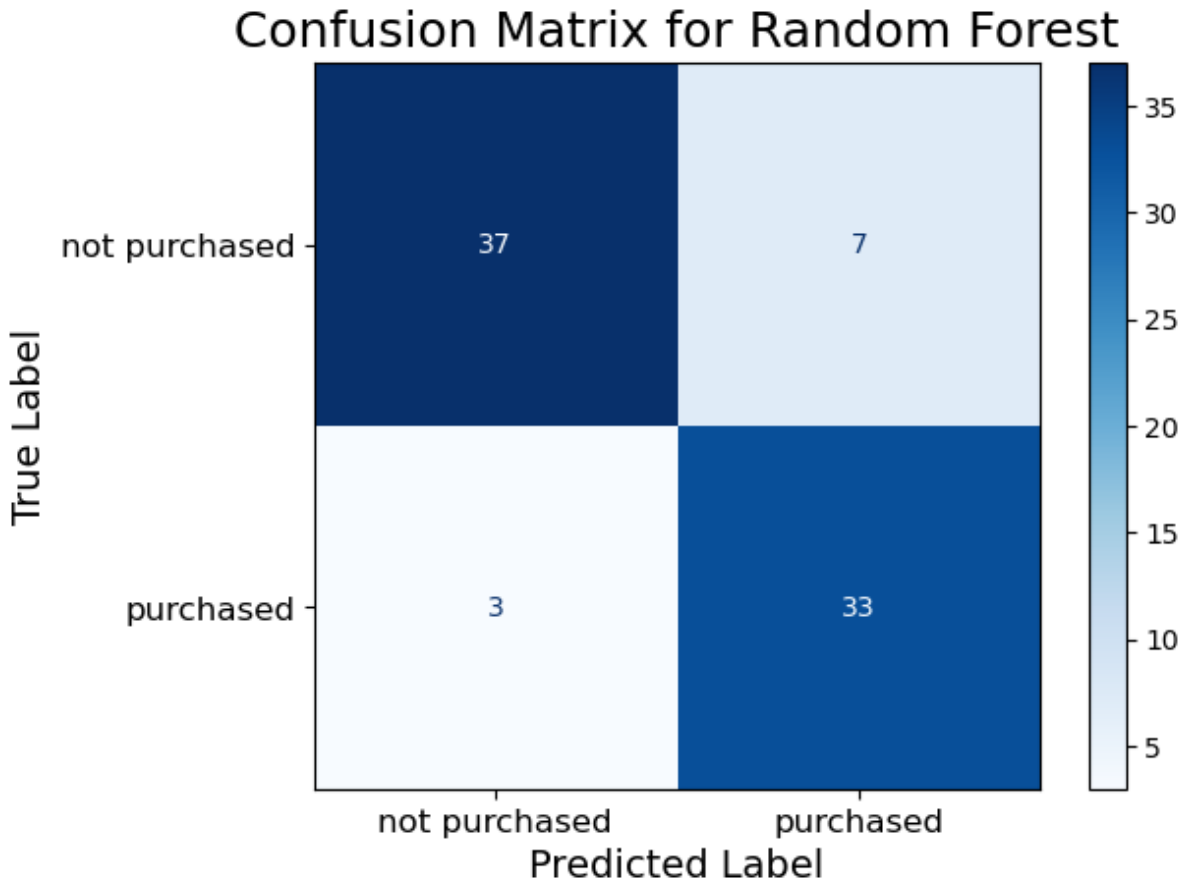
	학습용	테스트용
샘플 개수	320	80
Purchased = 0	213	44
Purchased = 1	107	36

최적의 하이퍼파라미터를 기반으로 구현된 모델에 학습용 데이터를 학습시켰습니다.

다음으로, 최종 모델의 성능을 평가하기 위해 학습 데이터와 테스트 데이터에서의 정확도를 계산했습니다.

- **학습 데이터 정확도**: 0.940625
- **최종 테스트 데이터 정확도**: 0.875

그리고, 최종 테스트 데이터에 대한 예측 결과와 실제 값을 비교하기 위해 혼동 행렬을 시각화했습니다.

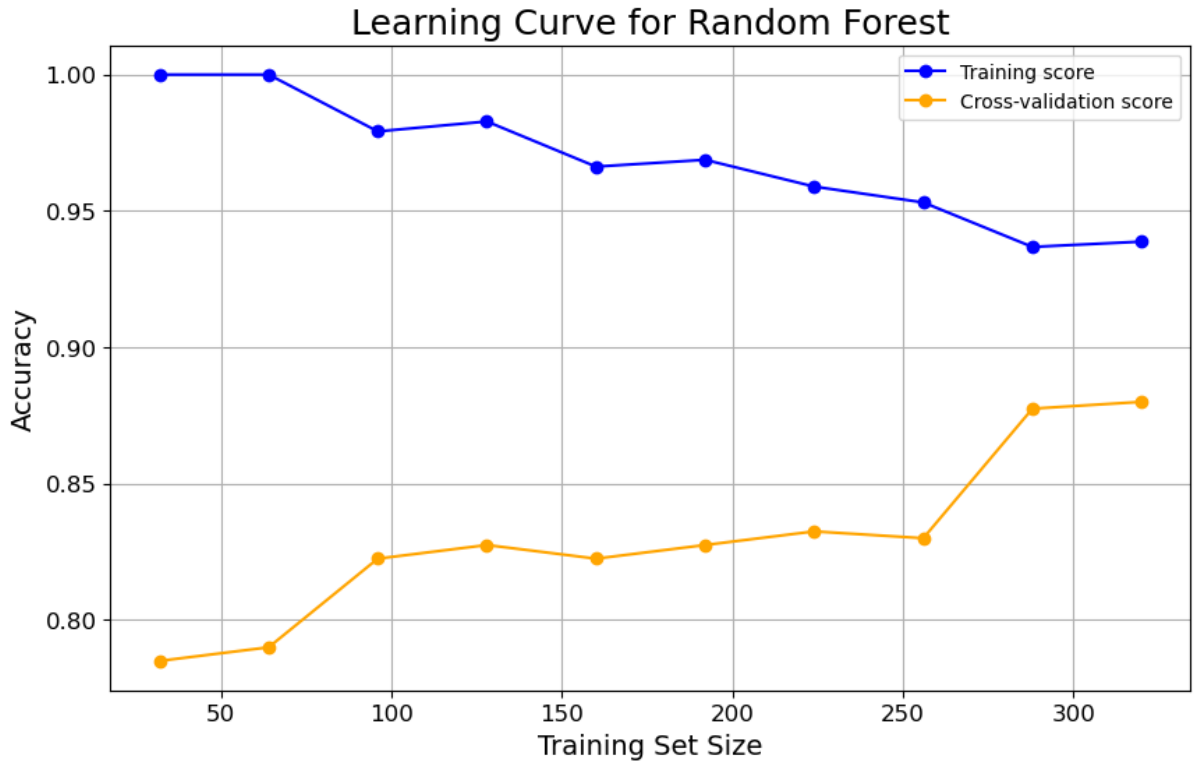


최종 테스트 데이터의 혼동 행렬을 보면, 모델은 7:1 (정확도 87.5%)의 예측 성공률을 보였습니다.

(이 정확도는 GridSearchCV를 통한 교차 검증 정확도와는 별개로, 최종 모델이 이전에 보지 않은 테스트 데이터에서 평가된 성능입니다.)

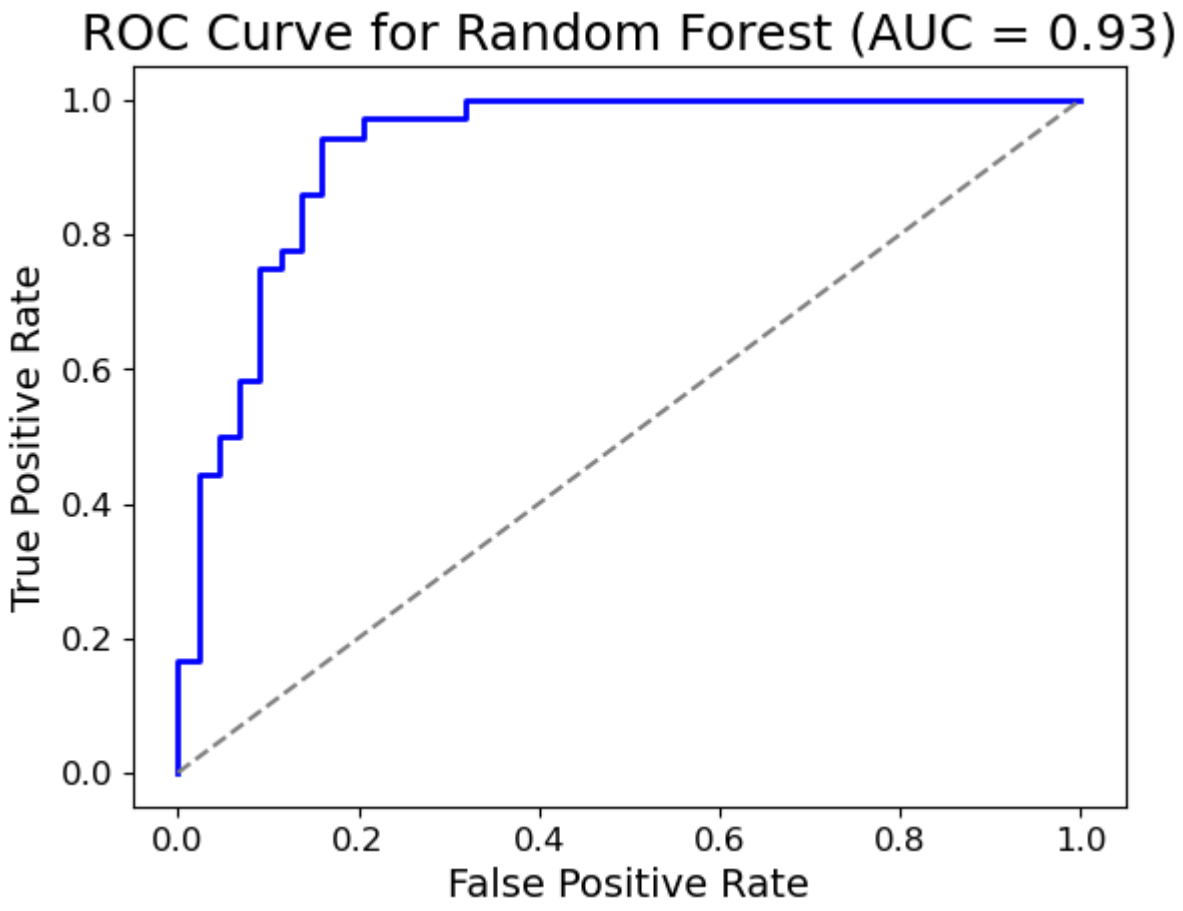
3. 모델 검증

모델의 학습 추이를 보기 위해 학습 곡선을 그려 보았습니다.



정확도와 학습 곡선을 보면, 학습 데이터의 정확도는 100%에서 94%까지 감소하고, 검증 데이터의 정확도는 80% 미만부터 87% 까지 증가하는 양상을 보아, 데이터를 학습할수록 모델 성능이 향상되는 것으로 보여집니다.

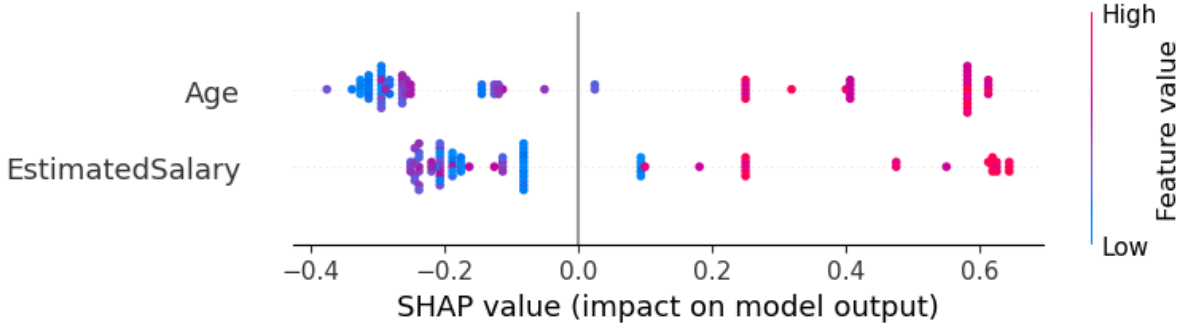
그리고 ROC 커브도 시각화해보았습니다.



ROC 커브가 왼쪽 상단으로 볼록하며, AUC(ROC 아래의 면적)가 0.93으로 1에 가까운 것을 통해 모델이 우수한 성능을 보였음을 확인할 수 있었습니다.

그리고 SHAP를 이용해 설명 가능한 인공지능(XAI) 기법으로 feature들의 영향력도 시각화 해보았습니다.

(feature 0 : 'Age', feature 1 : 'EstimatedSalary')



그래프를 보면,

- **Age**: 값이 클수록(빨간색) 모델 출력에 양의 방향으로 영향을 미치고 있습니다.
- **EstimatedSalary**: 대부분의 값이 모델에 미치는 영향이 작거나 음의 방향(특히 낮은 값일 때, 파란색)으로 나타납니다.

전체적으로 'Age'가 "EstimatedSalary'보다 영향력이 강한 것을 알 수 있었습니다.

결론

이번 실습에서는 SNS상의 광고를 시청한 사용자들의 나이, 예상 소득, 광고 상품 구매 여부가 담긴 데이터를 가지고 나이와 예상 소득을 feature로 설정하고, 광고 상품 구매 여부를 target으로 설정하여 사용자의 상품 구매 여부를 예측해 보는 모델을 구현했습니다. feature들에 표준화 과정을 거친 후에, 모델 선정에는 Stratified k-Fold 교차 검증을 사용하여 최적의 모델을 선택하고, 최적의 파라미터는 gridsearch 교차 검증으로 선별하여 최적의 모델이 나오도록 했습니다. 모델의 최종 테스트 결과, 혼동 행렬과 정확도를 계산하니 87.5%의 정확도로 아주 높다고 보이지는 않지만, 학습 곡선의 추이와 roc 커브 등의 지표에서 볼 수 있듯이 모델의 성능이 우수함을 알 수 있었습니다. SHAP를 이용해 XAI 기법으로 feature들의 영향력을 확인해 보니, 사용자의 나이가 예상 소득보다 영향력이 큰 것을 알 수 있었습니다.

이 모델을 통해, SNS 사용자들의 나이와 예상 소득만으로도 광고 상품 구매 여부를 예측할 수 있음을 확인했습니다. 이를 비즈니스적인 측면에서 본다면, 나이와 예상 소득으로 예측값을 산출하고, 광고가 사용자에게 적합한 광고인지 판별하여 사용자에게 필요한 상품을 제공할 확률을 높이는 데 이용할 수 있을 것으로 보입니다.

개선점

- SNS 사용자들의 데이터이므로, SNS 사용자들이 어떤 게시물을 자주 보는지, 어떤 게시물을 업로드했는지, 어떤 키워드를 검색하는지 등의 추가 데이터가 있다면 그에 따라 맞춤 광고의 제품군도 분류하여 더 퀄리티가 높은 광고를 노출시킬 수 있을 것입니다.

소스 코드

보고서 제출 시에 같이 첨부하겠습니다.

파일 이름 : 8.30_실습.ipynb

데이터셋 'Social_Network_Ads.csv'를 소스 코드 파일과 같은 폴더에 넣고 소스 코드를 실행시키시면 됩니다. (데이터셋 파일 경로는 ./Social_Network_Ads.csv)

참고 문헌

수업에서 배운 내용을 토대로 실습했습니다.