

고객 구매 데이터 기반 성별 예측 모델링 보고서

본 보고서는 백화점 고객 구매 데이터를 활용하여 고객의 성별을 예측하는 분류 모델 구축 및 평가 과정을 담고 있습니다.

프로젝트 목적

- 고객의 구매 행동 데이터(X.csv, y.csv)를 활용하여 성별을 효과적으로 예측하는 머신러닝 모델 구축
- 다양한 알고리즘(Logistic Regression, Decision Tree, Random Forest, XGBoost)의 성능 비교 및 최적의 모델 선정

데이터 분석 및 전처리 과정

1 데이터 개요

- X 데이터:** 고객 3,500명의 구매 이력 및 행동 데이터 (독립변수)
- y 데이터:** 고객의 성별 (종속변수, binary classification: 남성, 여성)

2 데이터 탐색 및 처리

- 결측치 및 이상치 점검
 - 범주형 변수 처리 (pd.get_dummies 활용하여 원-핫 인코딩 적용)
 - 데이터 표준화(StandardScaler) 적용하여 모델 성능 최적화
 - 데이터 분할 (train_test_split): 학습 데이터 80%, 테스트 데이터 20%
-

모델링 및 평가 결과

아래 4개의 모델로 성별 예측을 수행하였으며, 각 모델의 성능을 정확도(Accuracy) 기준으로 평가하였습니다.


분류 모델	정확도 (Accuracy)
-------	----------------

Logistic Regression	0.65
---------------------	------

Decision Tree	0.58
---------------	------

Random Forest	0.67
---------------	------

XGBoost	0.68
---------	-------------

 **최적의 모델: XGBoost**가 가장 높은 성능을 기록했습니다.