기계학습 캐글 데이터 실습 보고서

KU KONKUK UNIVERSITY 건국대학교 정보통신대학원

제출자 : 인공지능 전공 김민재

건국대학교 정보통신대학원

기계학습 캐글 데이터 실습 보고서

사용한 캐글 데이터셋과 분석 내용

사용 데이터 셋 : https://www.kaggle.com/datasets/bhadramohit/customer-shopping-latest-trends-dataset

위의 데이터 셋을 사용하여

다음과 같은 분석을 진행해보려고 합니다.

1. 인구 통계 : 나이 분포, 성별 비율, 주요 위치 분석

2. 구매 트렌드 : 계절별로 인기 상품, 가장 많이 구매된 카테고리와 색상 분석

3. 구매 성향 : 평균 구매 금액, 평점, 할인 적용 프로모션 코드 사용 여부 분석

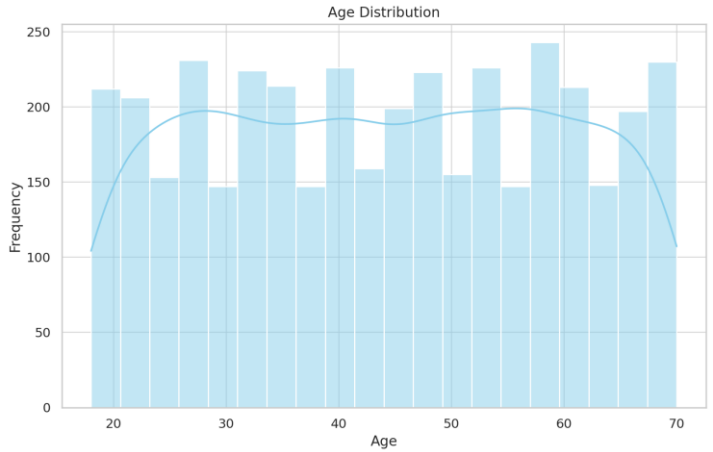
4. 지불 및 배송 : 선호 결제 방법 및 배송 유형 확인

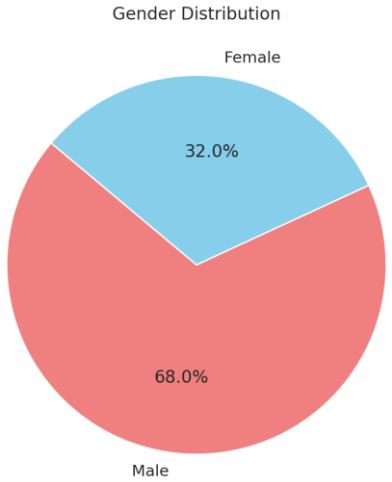
5. 고객 행동 : 구매 빈도 및 이전 구매 기록 기반으로 고객 세분화

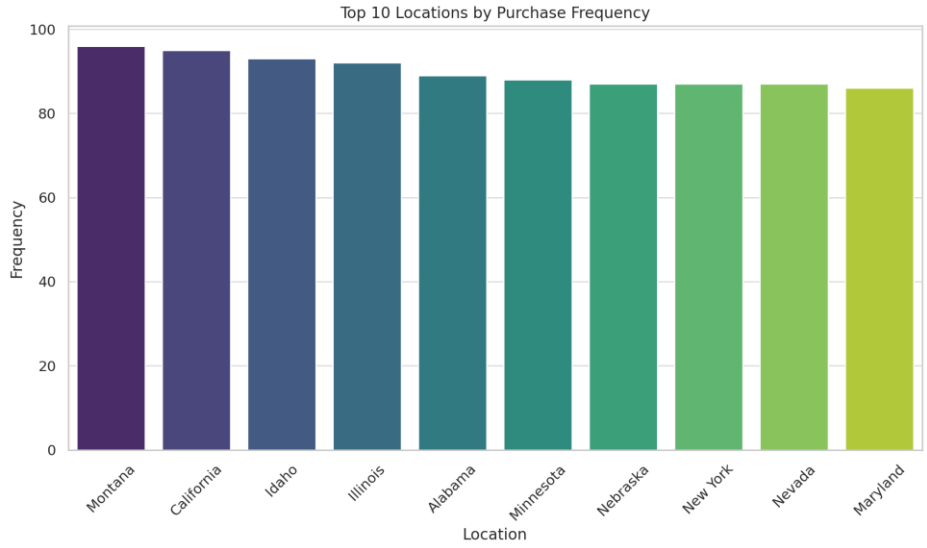
건국대학교 정보통신대학원

기계학습 캐글 데이터 실습 보고서

데이터셋의 시각화

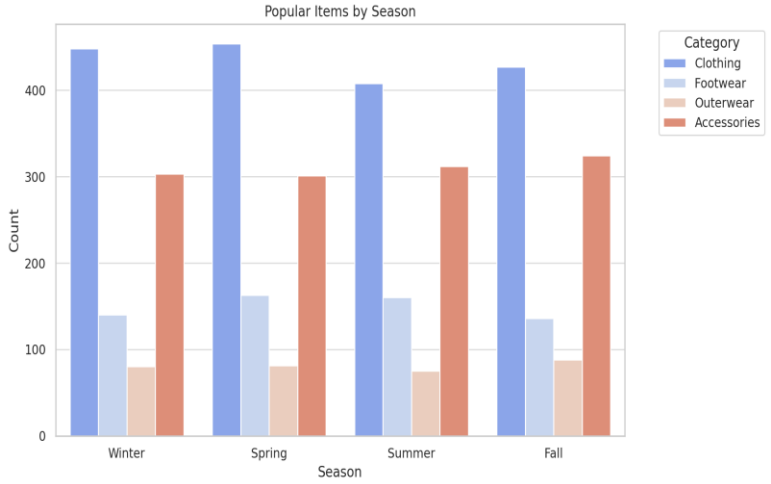


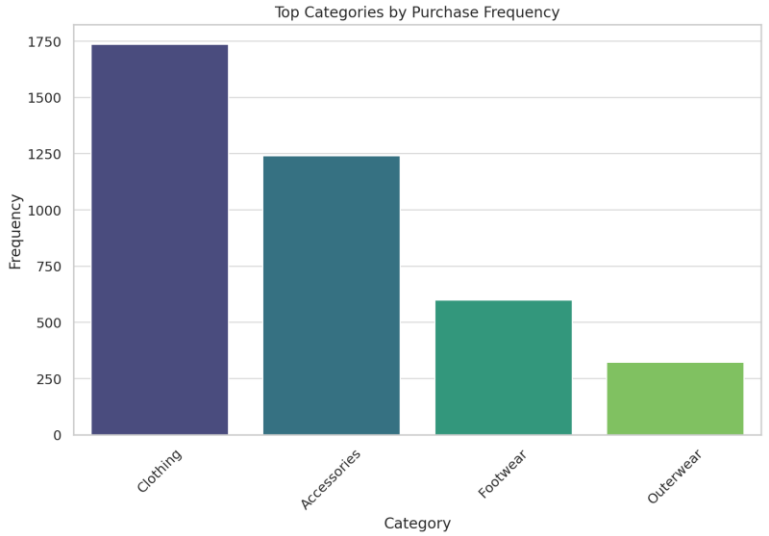


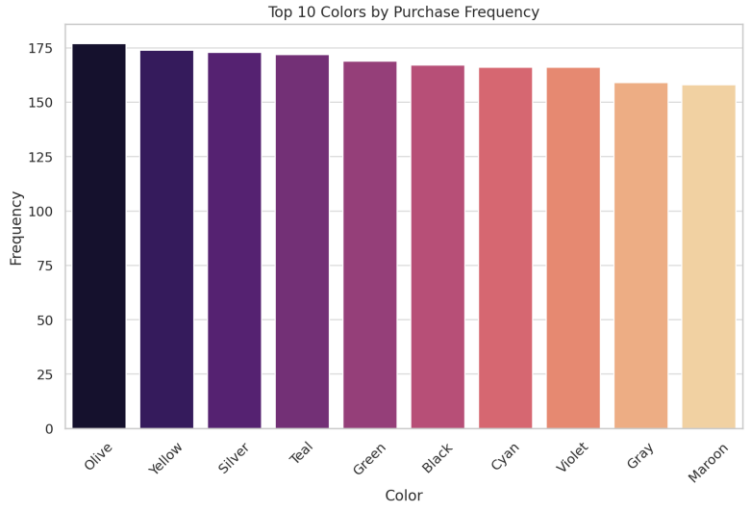


1. 인구 통계 : 나이 분포, 성별 비율, 주요 위치 분석

데이터셋의 시각화

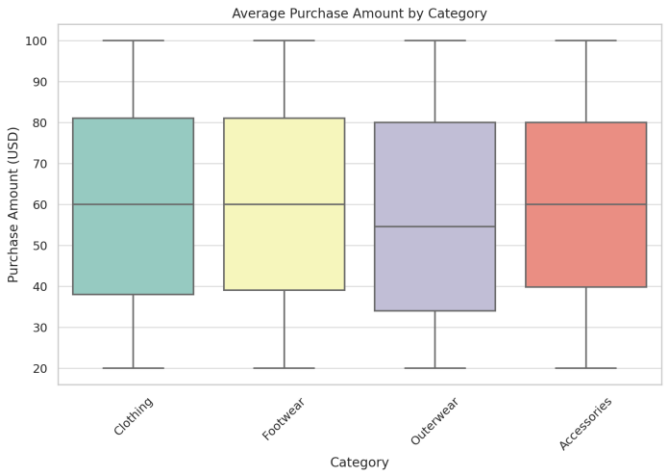


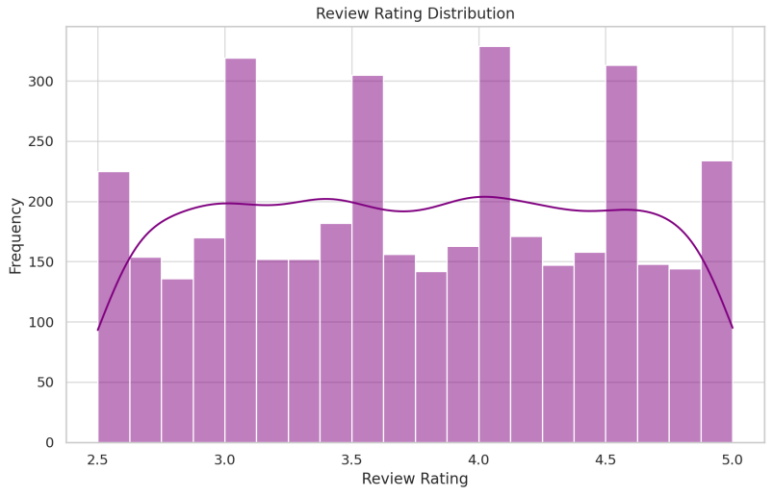




2. 구매 트렌드 : 계절별로 인기 상품, 가장 많이 구매된 카테고리와 색상 분석

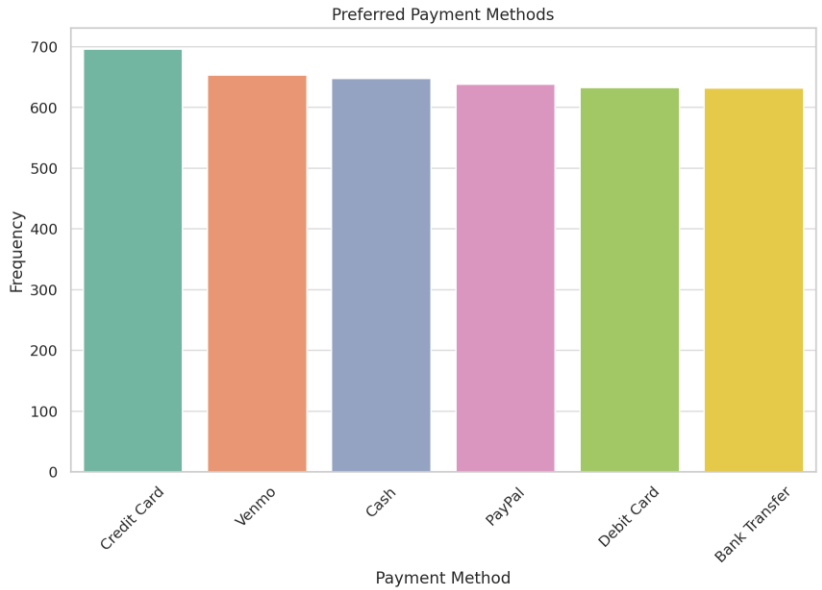
데이터셋의 시각화





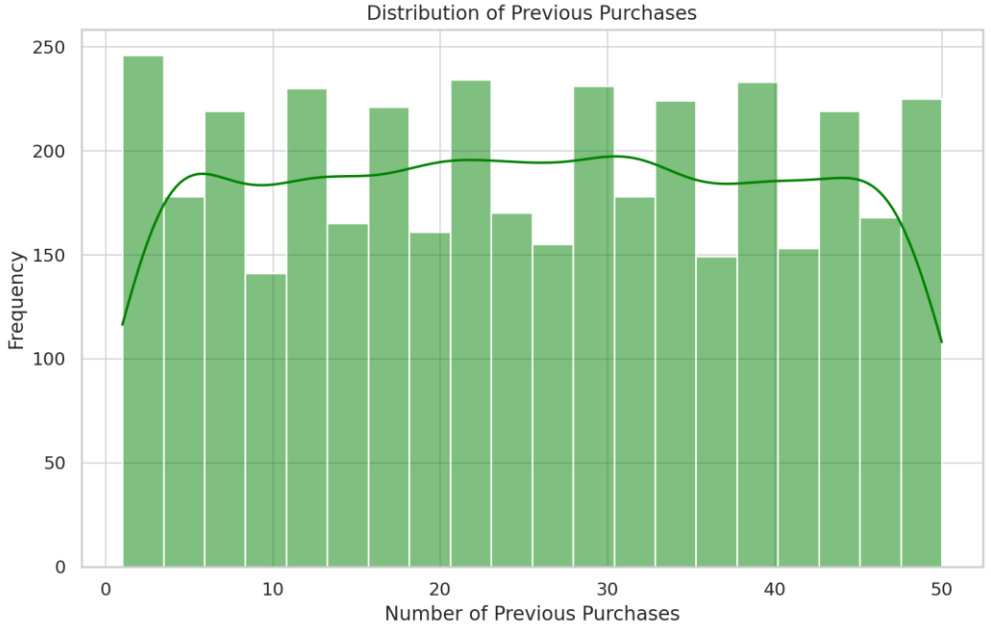
3. 구매 성향 : 평균 구매 금액, 평점, 할인 적용 프로모션 코드 사용 여부 분석

데이터셋의 시각화



4. 지불 및 배송 : 선호 결제 방법 및 배송 유형 확인

데이터셋의 시각화



5. 고객 행동 : 구매 빈도 및 이전 구매 기록 기반으로 고객 세분화

ML 모델 선정

모델을 선정하기 위하여 5개 모두 분석하는 것이 아니라 그 중 2가지를 선택하여 ML을 진행하였습니다.

1. 분류 모델

- 할인 적용 여부와 프로모션 코드 사용 여부 예측

사용 모델 : 로지스틱 회귀 (Logistic Regression) / XGBoost

2. 회귀 모델

- 구매 금액과 리뷰 평점 예츨

사용 모델 : 선형 회귀 (Linear Regression) / XGBoost

건국대학교 정보통신대학원

기계학습 캐글 데이터 실습 보고서

학습 실행 및 결과

```
Logistic Regression Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       422
           1       1.00      1.00      1.00       358

    accuracy                           1.00       780
   macro avg       1.00      1.00      1.00       780
weighted avg       1.00      1.00      1.00       780

XGBoost Classification Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       422
           1       1.00      1.00      1.00       358

    accuracy                           1.00       780
   macro avg       1.00      1.00      1.00       780
weighted avg       1.00      1.00      1.00       780

/usr/local/lib/python3.10/dist-packages/xgboost/core.py:158: UserWarning: [10:51:02] WARNING: /workspace/src/learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
Linear Regression MSE: 562.5689377435878
Linear Regression R2 Score: -0.005337133725525467
XGBoost Regression MSE: 717.0823891714352
XGBoost Regression R2 Score: -0.2814599275588989
```

분류 모델 (Logistic Regression, XGBoost)

- **Logistic Regression**:
  - **정확도 (Accuracy)**: 1.0
  - **정밀도 (Precision)**, **재현율 (Recall)**, **F1 점수** 모두 완벽한 성능을 보임.
  - 데이터셋이 매우 균일하거나, 특정 특성이 분류에 강력한 영향을 미쳤을 가능성이 있음.
- **XGBoost Classification**:
  - Logistic Regression과 동일한 완벽한 성능.
  - **정확도**: 1.0
  - 동일한 경고 메시지가 출력되었으나 학습에 영향은 없음.

회귀 모델 (Linear Regression, XGBoost Regression)

- **Linear Regression**:
  - **평균 제곱 오차 (MSE)**: 562.57
  - **결정 계수 (R^2)**: -0.0053 (모델이 데이터 변동성을 거의 설명하지 못함)
- **XGBoost Regression**:
  - **평균 제곱 오차 (MSE)**: 717.08
  - **결정 계수 (R^2)**: -0.2815 (예측 성능이 더 나쁨)

건국대학교 정보통신대학원

기계학습 캐글 데이터 실습 보고서

학습 실행 및 결과

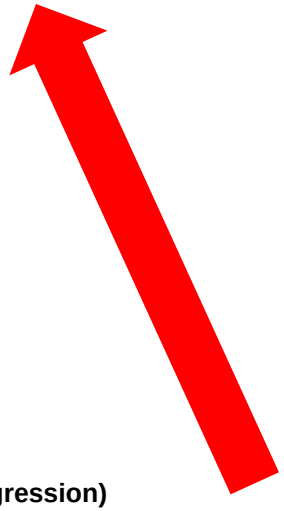
```
Logistic Regression Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       422
           1       1.00      1.00      1.00       358

    accuracy                           1.00       780
   macro avg       1.00      1.00      1.00       780
weighted avg       1.00      1.00      1.00       780

XGBoost Classification Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       422
           1       1.00      1.00      1.00       358

    accuracy                           1.00       780
   macro avg       1.00      1.00      1.00       780
weighted avg       1.00      1.00      1.00       780

/usr/local/lib/python3.10/dist-packages/xgboost/core.py:158: UserWarning: [10:51:02] WARNING: /workspace/src/learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
Linear Regression MSE: 562.5689377435878
Linear Regression R2 Score: -0.005337133725525467
XGBoost Regression MSE: 717.0823891714352
XGBoost Regression R2 Score: -0.2814599275588989
```

분류 모델 (Logistic Regression, XGBoost)

- **Logistic Regression**:
  - **정확도 (Accuracy)**: 1.0
  - **정밀도 (Precision)**, **재현율 (Recall)**, **F1 점수** 모두 완벽한 성능을 보임.
  - 데이터셋이 매우 균일하거나, 특정 특성이 분류에 강력한 영향을 미쳤을 가능성이 있음.
- **XGBoost Classification**:
  - Logistic Regression과 동일한 완벽한 성능.
  - **정확도**: 1.0
  - 동일한 경고 메시지가 출력되었으나 학습에 영향은 없음.

교차 검증 (Cross-Validation):

- K-Fold 교차 검증을 통해 모델 성능을 안정적으로 평가.



회귀 모델 (Linear Regression, XGBoost Regression)

- **Linear Regression**:
  - **평균 제곱 오차 (MSE)**: 562.57
  - **결정 계수 (R^2)**: -0.0053 (모델이 데이터 변동성을 거의 설명하지 못함)
- **XGBoost Regression**:
  - **평균 제곱 오차 (MSE)**: 717.08
  - **결정 계수 (R^2)**: -0.2815 (예측 성능이 더 나쁨)

건국대학교 정보통신대학원

기계학습 캐글 데이터 실습 보고서

재학습 및 실행 결과

```
Linear Regression Cross-Validation R2 Scores: [-0.00533713 -0.00504004 -0.00109986 -0.00494228 -0.00558807]
Linear Regression Average R2 Score: -0.004401476831551632
XGBoost Regression Cross-Validation R2 Scores: [-0.00679505 -0.01849806 -0.01811576 -0.01142967 -0.02763844]
XGBoost Regression Average R2 Score: -0.016495394706726074
```

Linear Regression 결과
- **R^2 점수 (Fold 별)**: [-0.0053, -0.0050, -0.0011, -0.0049, -0.0056]
- **평균 R^2 점수**: -0.0044

분석:
- R^2 점수가 음수라는 것은 모델이 데이터의 변동성을 거의 설명하지 못한다는 의미
- 기본 선형 회귀 모델은 비선형적 패턴을 처리할 수 없기 때문에, 변수 간 상관관계가 낮은 경우 성능이 제한적

XGBoost Regression 결과
- **R^2 점수 (Fold 별)**: [-0.0068, -0.0185, -0.0181, -0.0114, -0.0276]
- **평균 R^2 점수**: -0.0165

분석:
- XGBoost 회귀 모델 또한 평균 R^2 점수가 음수로 나타나, 데이터의 변동성을 잘 설명하지 못하고 있음
- 이는 학습률과 반복 수 등을 튜닝해도 데이터 특성 자체가 모델 학습에 적합하지 않을 가능성을 시사

비선형적 패턴 존재 가능성이 있어
랜덤 포레스트 회귀(Random Forest Regressor)를 사용하여 비선형 관계를 더 잘 학습하도록 시도해 보았습니다

건국대학교 정보통신대학원

기계학습 캐글 데이터 실습 보고서

재학습 및 실행 결과

```
Random Forest Regression Cross-Validation R2 Scores: [-0.16052159 -0.15442018 -0.09380611 -0.15407574 -0.15191051]
Random Forest Regression Average R2 Score: -0.14294682801819136
```

랜덤 포레스트 회귀 결과
- **R^2 점수 (Fold 별)**: [-0.1605, -0.1544, -0.0938, -0.1541, -0.1519]
- **평균 R^2 점수**: -0.1429

```
Test Loss (MSE): 569.013427734375
```

분석

1.음수 R^2 점수:
- 모델이 입력 변수와 출력 변수 간의 변동성을 잘 설명하지 못하고 있음을 나타냄
- 랜덤 포레스트 모델임에도 음수의 R^2 점수가 나온 것은 데이터 자체에 주요 정보가 부족하거나, 목표 변수(Purchase Amount (USD))와 입력 변수 간 상관관계가 매우 낮을 가능성이 큼

2.입력 변수의 한계:
- 현재 사용 중인 변수(Age, Gender, Category, Season, Previous Purchases, Promo Code Used)가
- Purchase Amount (USD)를 설명하기에 충분하지 않을 수 있음

마지막으로 데이터 노이즈를 제거하고
비선형 모델을 다룰때 유리한 신경망 모델을 사용하여 재학습 시키도록해보겠습니다.

재학습 및 실행 결과

```
Test Loss (MSE): 569.013427734375
```

신경망 모델 결과

- **Test Loss (MSE)**: 569.01
- 이 값은 기존 **랜덤 포레스트**나 **XGBoost** 모델과 비슷한 수준으로, 여전히 높은 편

최종 분석

분류 모델

모델 사용 : Logistic Regression, XGBoost Classification

결과 :
- 정확도 : 100프로정도의 완벽한 분류 결과를 보임

- 정밀도, 재현율도 1.0

- 잠재적 과적합이 있을 수도 있음.

회귀 모델 (신경망)

모델 사용 : 다층 퍼셉트론(MLP, 신경망 모델)

결과 :

- 모델이 출력 변수(Purchase Amount (USD))의 변동성을 설명하지 못하며, 기존 랜덤 포레스트 및 XGBoost 모델과 유사한 수준의 성능을 보임

- 한계 분석 : 이상치 제거에도 데이터에 내재된 정보가 부족하거나 중요한 사항이 누락되었을 가능성 있음.

해결법 :
- 추가적인 변수 수집 / 비선형적 관계를 좀 더 잘 반영하여 데이터 품질 상향