kaggle 데이터를 활용한 머신러닝 예측

인공지능 전공 202476720
이정호

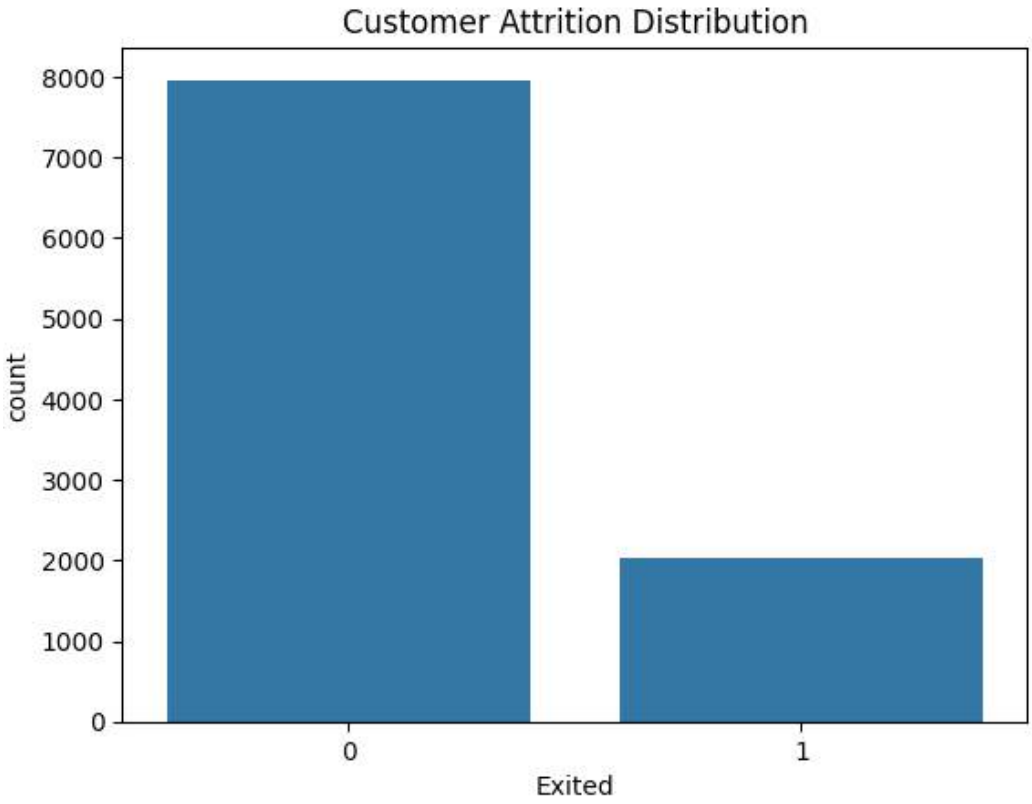
개요

본 실습에서는 BankCustomer.csv 데이터를 활용하여 고객 이탈 여부(Exited)를 예측하는 기계학습 모델의 성능을 비교하였다.
Logistic Regression(L1, L2), SVM 모델을 사용하여 학습을 수행하고, 다양한 시각화 분석을 통해 데이터 특성을 탐색하였다.

데이터 탐색 → 전처리 → 시각화 → 모델링 → 평가 → 혼동행렬 분석 순으로 진행하였다.

1. 데이터 로드 및 기본 정보 확인
kaggle에서 다운받은 BankCustomer.csv 파일을 불러와 기본 구조 확인
info(), describe()를 통해 결측치 및 변수 타입 확인
모든 변수에서 결측치 없음
타켓: Exited

2. 타깃 변수 분포 확인

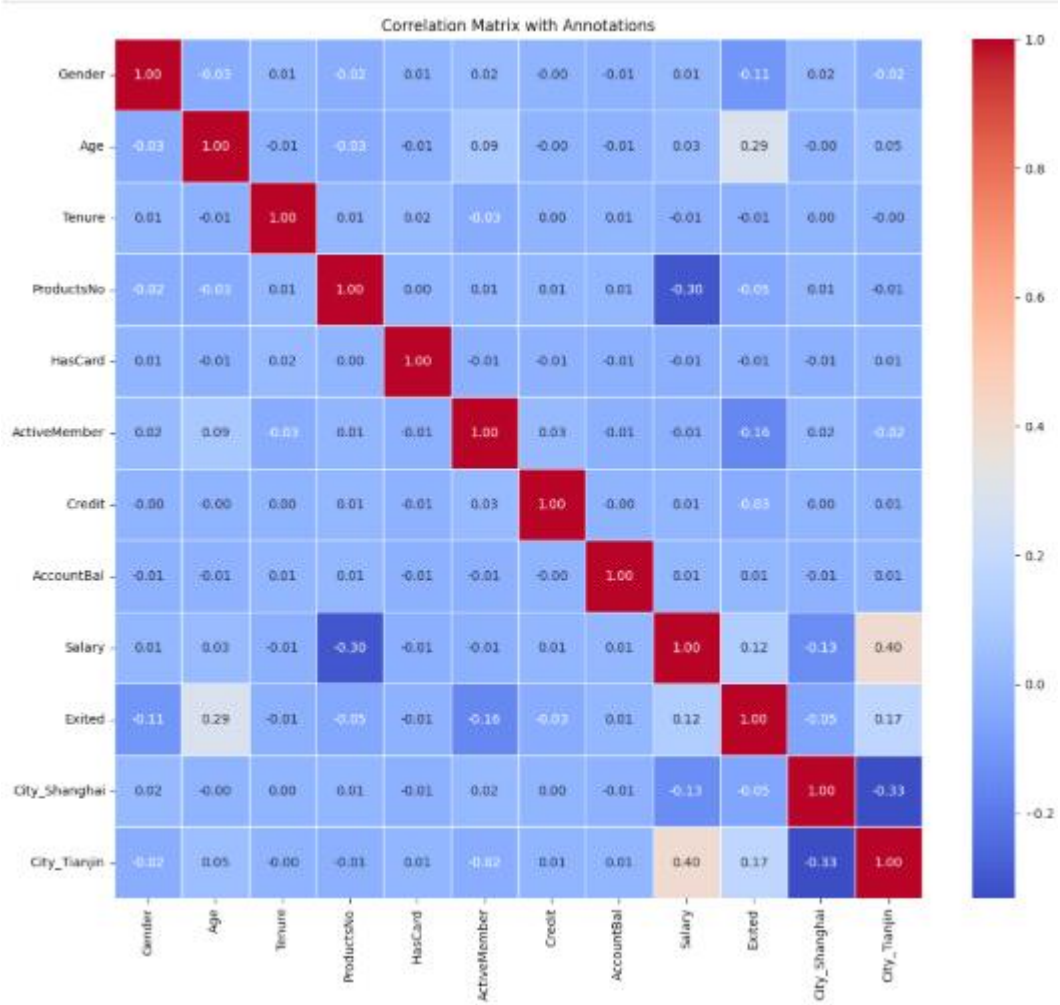


클래스 불균형 존재, 모델 학습 시 영향을 줄 수 있음.

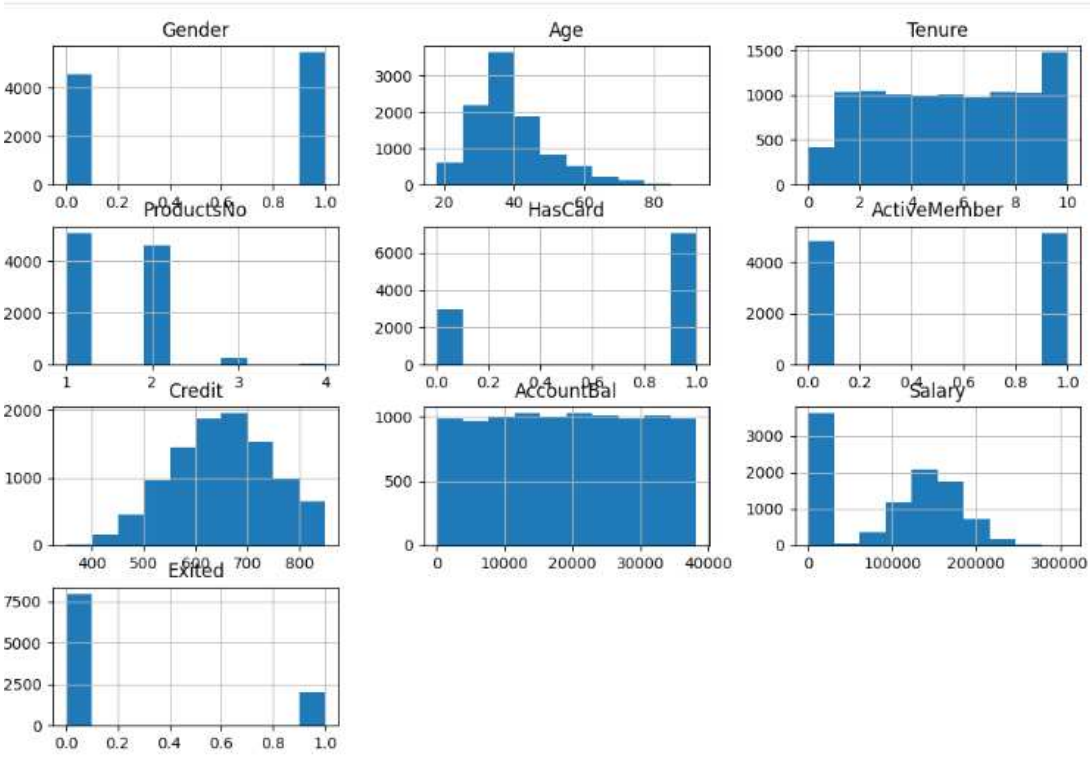
3. 데이터 전처리

불필요한 변수 제거 및 범주형 변수 인코딩

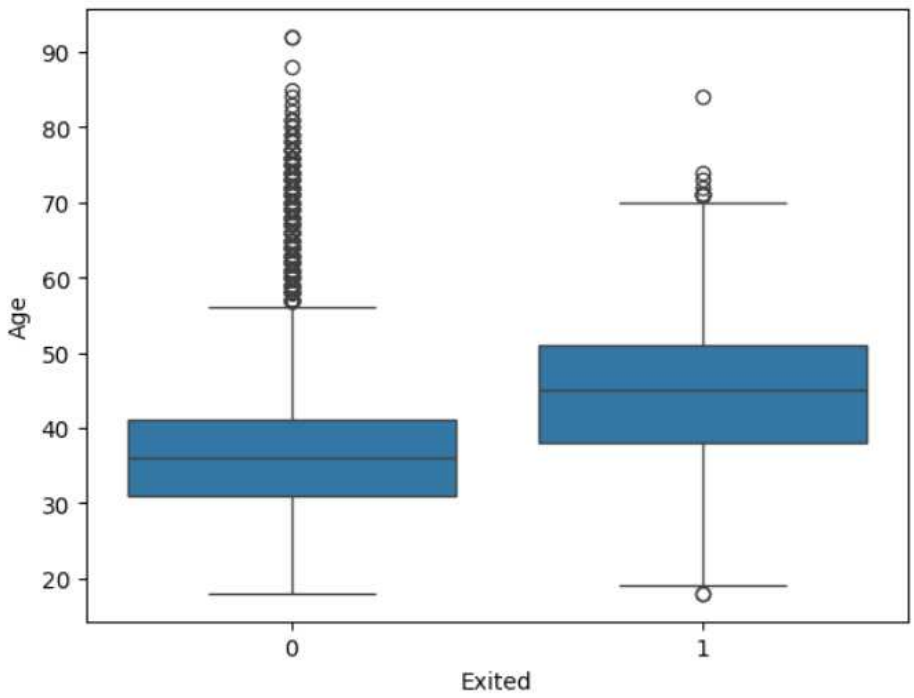
4. 데이터 시각화



<히트맵>



<각 칼럼 별 수치>



<연령별 이탈 정도>

5. 데이터 전처리 및 데이터 분할

모델 학습을 위해 데이터를 전처리 후 훈련 데이터와 테스트 데이터를 8:2 비율로 분할.

6.모델 학습

Logistic Regression (L1, L2)

- penalty='l1' (liblinear)
- penalty='l2' (lbfgs)

SVM

- 기본 커널 적용
- 분류용 SVC 모델 활용

7. 모델 평가

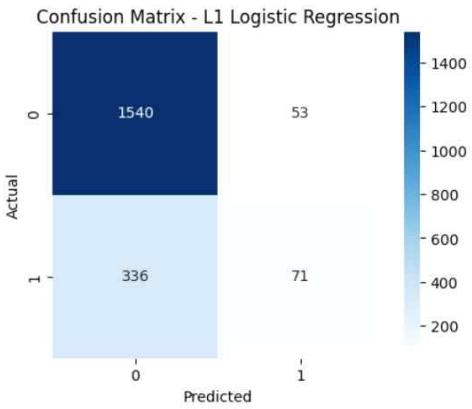
평가 지표: Accuracy

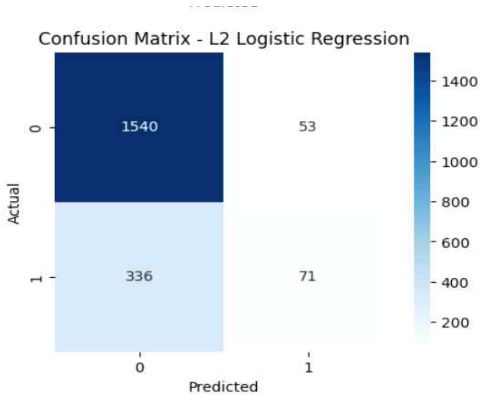
L1 Logistic Regression: 0.75

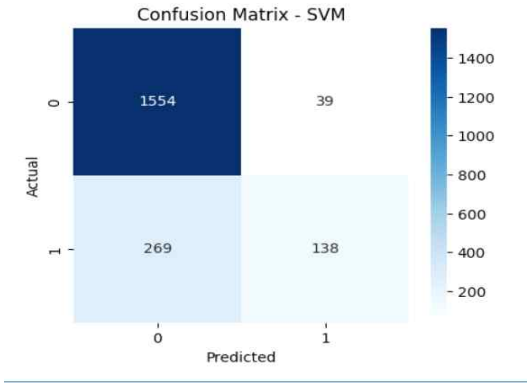
L2 Logistic Regression: 0.75

SVM: 0.81

8. Confusion Matrix







9. 최종 분석 결과

- Logistic Regression(L1/L2)은 규제 방식 차이로 인해 성능이 약간 다르게 나타났음

- SVM 모델이 상대적으로 안정적인 성능을 보였으나 모든 모델이 극적인 성능 차이는 없었음

- 고객 이탈 예측 문제에서는

클래스 불균형,

숫자/범주형 변수 조합,

상관관계 약함

등이 모델 성능에 영향을 준 것으로 판단됨.

최종적으로 SVM 또는 L2 Logistic Regression 모델이 가장 적합한 모델로 판단된다.