인공지능 데이터 전처리 결과 보고서

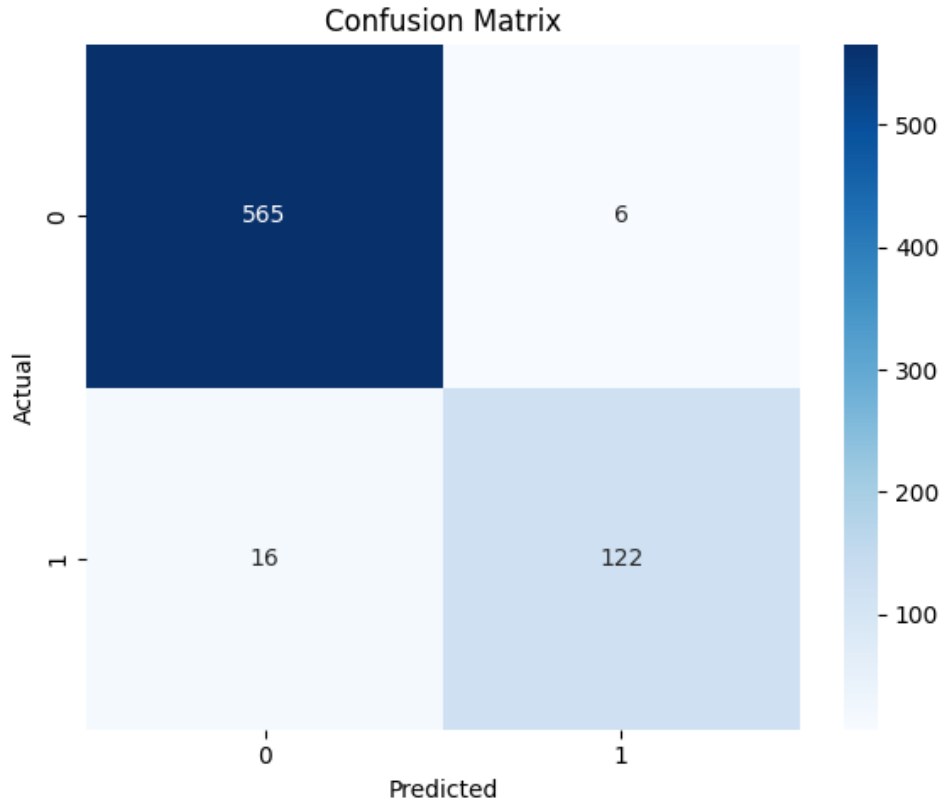
1. 수학 데이터

- 결측치: 수치형 컬럼 → 평균 번호 대체
- 이상치: IQR 기반 생성 및 제거
- 범주형: `get_dummies(drop_first=True)` 적용
- 클래스 불균형: SMOTE 적용 (Train only)
- 스케일링: `StandardScaler`

2. 평가

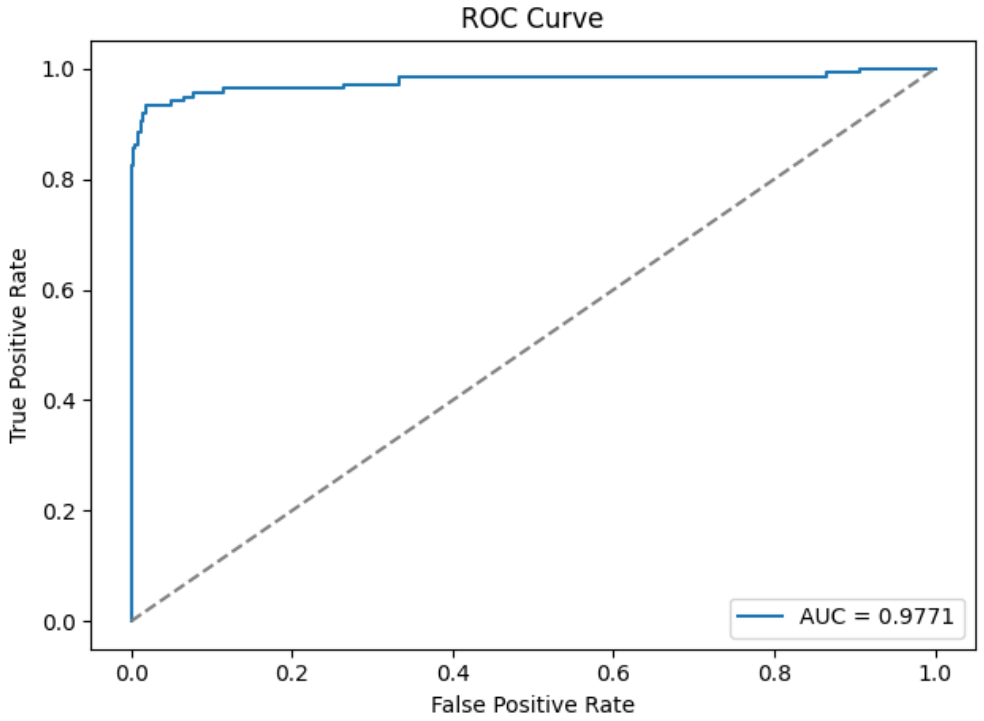
- 모델: XGBoostClassifier
- 학습 데이터: SMOTE + Scaling 적용 X_train_res
- Train Accuracy: 1.0000
- Test Accuracy: 0.9690
- Train F1 Score: 0.98+ (실제: 1.0000)
- Test F1 Score: 0.9173
- Precision: 0.9531 / Recall: 0.8841
- AUC Score: 0.9771
- 과적합 차이 (Train-Test F1): 0.0827
- 교차검증: CV 평균 F1 = 0.9600 / 표준편차 = 0.0432
- 분류 리포트:
- - Class 0: precision=0.97, recall=0.99, f1=0.98 (support=571)
- - Class 1: precision=0.95, recall=0.88, f1=0.92 (support=138)
- - Accuracy=0.97, Macro avg f1=0.95, Weighted avg f1=0.97

2. 평가

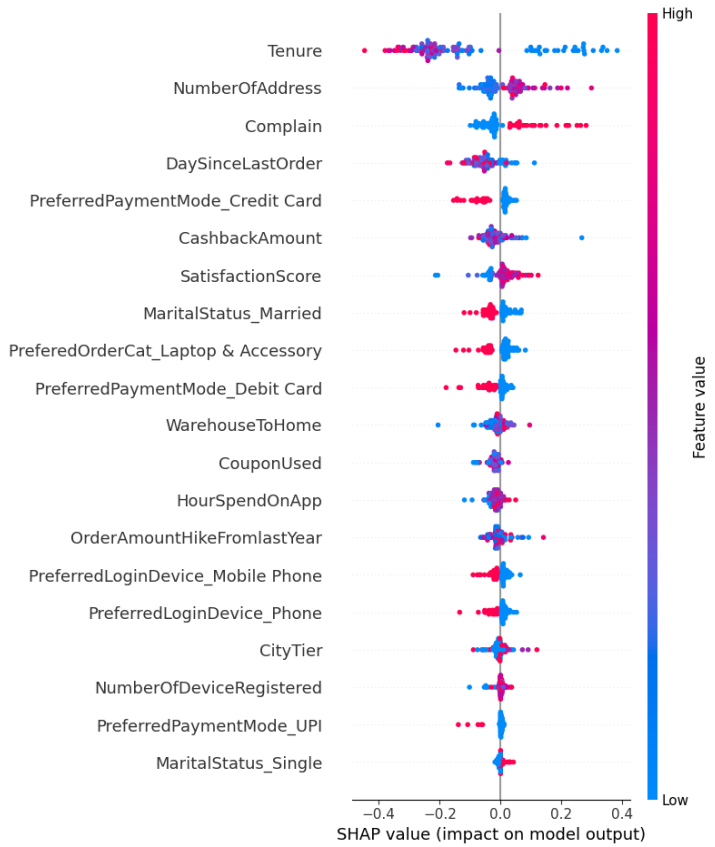


항목	값
True Negative (TN)	565
False Positive (FP)	6
False Negative (FN)	16
True Positive (TP)	122

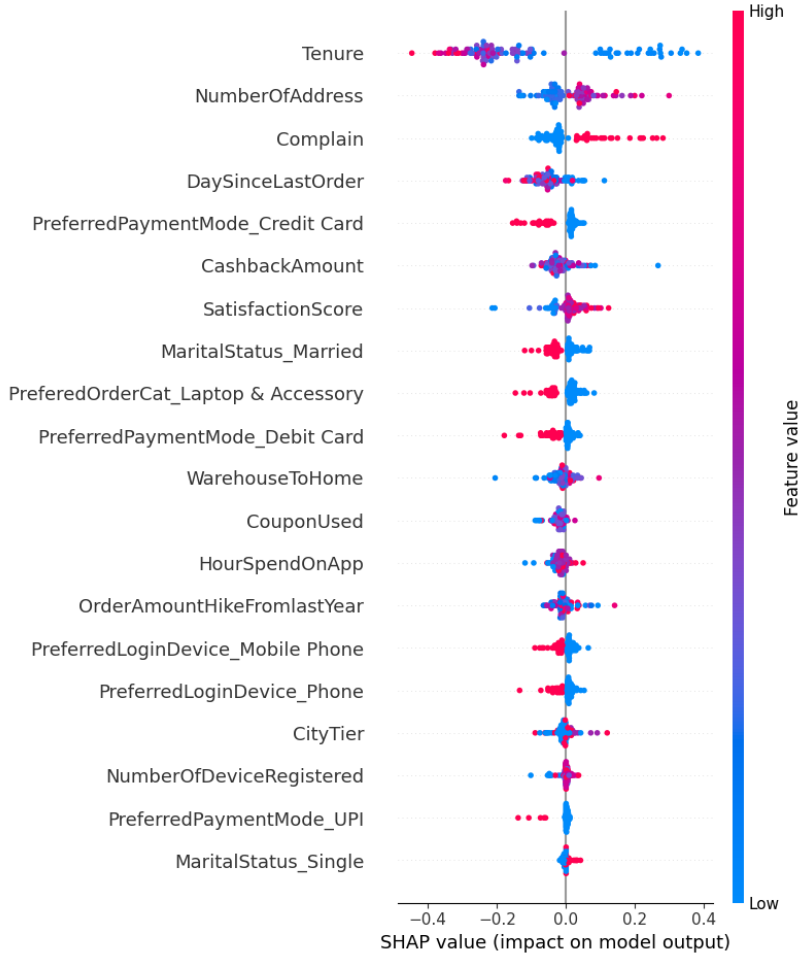
2. 평가



AUC = 0.9771 으로 매우뛰어난 성능을 보여주어
거의 완벽에 가까운 예측력을 보여줌



2. 평가



예측에 영향을 미친 주요 변수 (SHAP 분석)

변수명	영향도	요약
Tenure	최고 영향	이용 개월수가 길수록 (🔴) 이탈 확률 낮음
NumberOfAddress	매우 높음	주소 수가 많을수록 이탈 확률 높음
Complain	매우 높음	불만 제기 이력이 있는 고객은 이탈 가능성 높음
DaySinceLastOrder	높음	마지막 주문 이후 시간이 길수록 이탈 경향 증가
Preferred PaymentMode_Credit Card	높음	신용카드 선호자 중 일부 이탈 경향 존재
CashbackAmount	중간	캐시백 금액이 많거나 적을수록 예측에 혼합된 영향 존재
SatisfactionScore	중간	만족도가 낮을수록 이탈 확률 증가
MaritalStatus_Married	중간	기혼 상태가 이탈에 영향을 미침
PreferredDevice/Payment	낮음	디바이스나 결제 수단 관련 변수들은 보조적인 영향
MaritalStatus_Single	미미	일부 샘플에서는 낮은 영향력 가짐

3. 학습된 인공지능 모델

- 모델: XGBoostClassifier
- 저장: joblib 또는 pickle
- 파일: xgboost_best_model.pkl
- 입력 피처 수: 28개
- 입력 컬럼 목록 파일 (선택): model_features.json
- 사용 환경: Python 3.12+, scikit-learn 1.3+, xgboost 1.7+