기계학습 1조

PRESENTATION

기계학습 기반 고객 정기예금 가입 확률 예측 프로그램

22013365 김여경
20011752 임상수

목차

01 프로젝트 주제 및 목적

프로젝트의 주제

기계학습 기반 정기예금 가입 가능성 예측 모델 개발 및 마케팅 전략 적용

목표 : 고객 특성과 경제 지표를 바탕으로 정기예금 가입 가능성을 예측하는 모델 개발

활용 : 예측한 결과를 기반으로 마케팅 타켓 고객을 분류하여 ROI 개선 및 비용 효율화 실현

프로젝트의 목적

마케팅 자원 효율화
예측 모델을 통해 가입 가능성이 높은 고객을 우선 선정
→ 불필요한 연락 최소화

자동화 적용 가능성
예측 확률 기반 고객 분류
→ 콜센터,문자 캠페인 자동화에 직접 적용 가능

성과 향상 및 ROI 개선
최적 임계값 도출로 전환율 상승 및 마케팅 투자 대비 수익률 향상

02 프로젝트 분석

1	2	3	4
데이터 분석	**어떠한 문제점이 있었고 어떻게 해결했는가?**	**머신러닝 모델별 예측 성능 차이는 어느 정도이며, 어떤 모델을 어떠한 기준으로 최종 선택할 것인가?**	**예측 결과를 활용한 실제 마케팅 전략을 통해 기대할 수 있는 효과는?**
변수 분포와 변수간의 상관관계 분석 데이터셋 정보, 통계량 분석	결측치 처리, 클래스 불균형 문제, 하이퍼파라미터 탐색 시간 문제 등에 대한 해결 방안	불균형 데이터 특성상 F1 Score을 기준으로 선택	최우선 타겟 고객 우선 접촉, 마케팅 ROI 향상 기대

03 데이터 분석

데이터 셋 정보	Bank_Marketing_Dataset 전체 레코드 수: 41,188개 컬럼 수: 22개(연령, 직업, 교육, 혼인 상태, 대출 여부, 경제 지표 등)
결측치 정보	NaN형태의 결측치는 0개, 하지만 일부 범주형 컬럼(job, education, marital, poutcome 등)에 “unknown”이라는 문자열이 결측 정보가 기록되어 있어 전처리 과정에서 처리
타겟 분포	No : 36,548개 (88.73%) Yes : 4,640개 (11.27%) -> 클래스 불균형 존재

04 데이터 전처리 분석

결측치 처리

문제: 원본 데이터에서 일부 column(ex.job, education)에 “unknown”형태의 결측 정보가 카테고리로 기록됨

해결방안

수치형 변수

: 중앙값(median)으로 결측 대체 → 극단치 영향 최소화

```
numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
    ('scaler', StandardScaler())
])
```

범주형 변수

: 최빈값(mode)으로 결측 대체 → 분포 왜곡 최소화, One-Hot 인코딩

```
categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])
```

이상치 처리

문제: duration(통화 시간)의 일부 극단치(몇천초이상), pdays(마지막 캠페인 이후 경과 일수)가 -1로 기록(미접촉 의미)

해결방안

median imputation + StandardScaler

median 대체 후 StandardScaler를 통해 평균 0, 분산 1로 정규화 → 이상치 분포가 크게 완화됨

05 모델 학습 방법

1

데이터 분할

# 80:20 분할, 층화추출로 클래스 비율 유지
train_test_split(X, y, test_size=0.2, stratify=y)

Stratify 적용으로 클래스 비율 동일하게 유지

2

전처리 파이프라인

#수치형: 결측치→ 표준화
#범주형: 결측치→ 원핫
preprocessor = ColumnTransformer([...])

모든 데이터를 동일한 스케일로 변환

3

4개 모델 동시 학습

models = {
'Logistic Regression': class_weight='balanced',
'Decision Tree': class_weight='balanced',
'Random Forest': class_weight='balanced',
'Gradient Boosting': 기본 설정
}

4

성능 평가

F1 score 중심 평가(불균형 데이터 특성)
5개 지표 종합 비교
:Accuracy, Precision, Recall, F1, ROC-AUC

5

최적 모델 선택

Gradient Boosting 최종 선택
모든 평가 지표에서 최고 성능 달성

05 모델 학습 방법

하이퍼파라미터 튜닝

GridSearchCV 적용

3-fold Cross Validation
F1 Score 기준 최적화
병렬 처리 (n_jobs=1) 활용

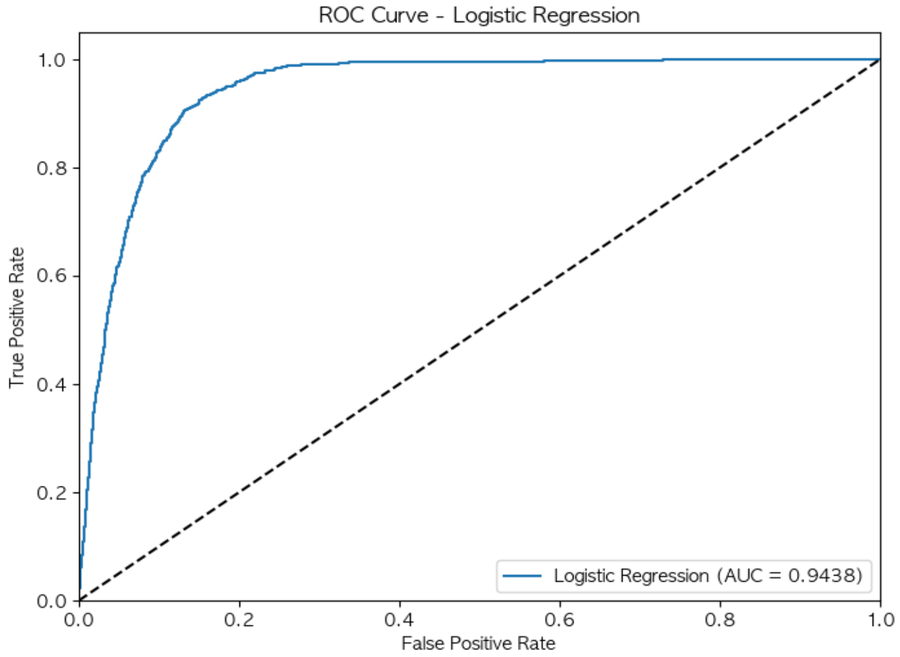
주요 튜닝 파라미터

```
param_grid = {
    'classifier__n_estimators': [100, 200],
    'classifier__max_depth': [None, 10, 20],
    'classifier__min_samples_split': [2, 5],
    'classifier__min_samples_leaf': [1, 2]
}
```

학습 방법의 특징

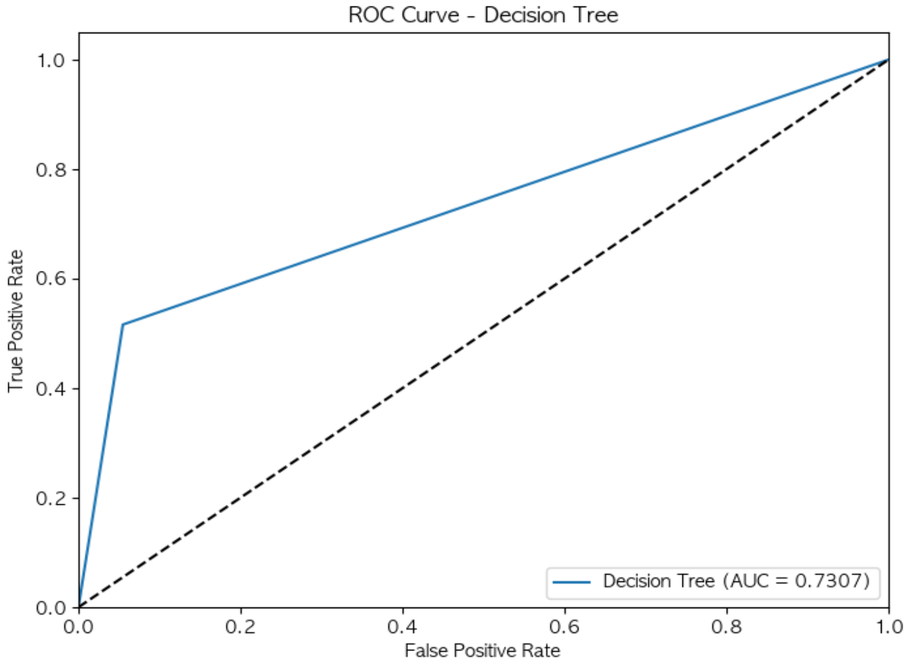
1. 파이프라인 구조: 전처리 모델 일체화로 실무 적용 용이
2. 클래스 균형: 불균형 문제 해결로 신뢰할만한 예측
3. 다중 모델 비교: 4개 알고리즘 동시 평가로 최적 선택
4. 체계적인 평가: 다양한 지표로 종합적인 성능 평가

05 모델 성능 평가 및 결과



Logistic Regression

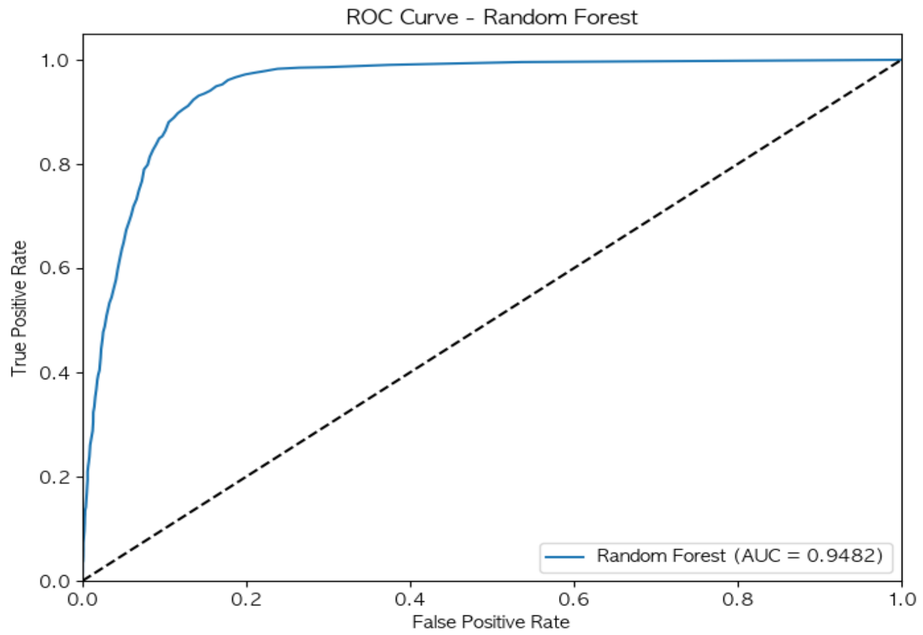
기본 선형 모델/ 해석 용이성



Decision Tree

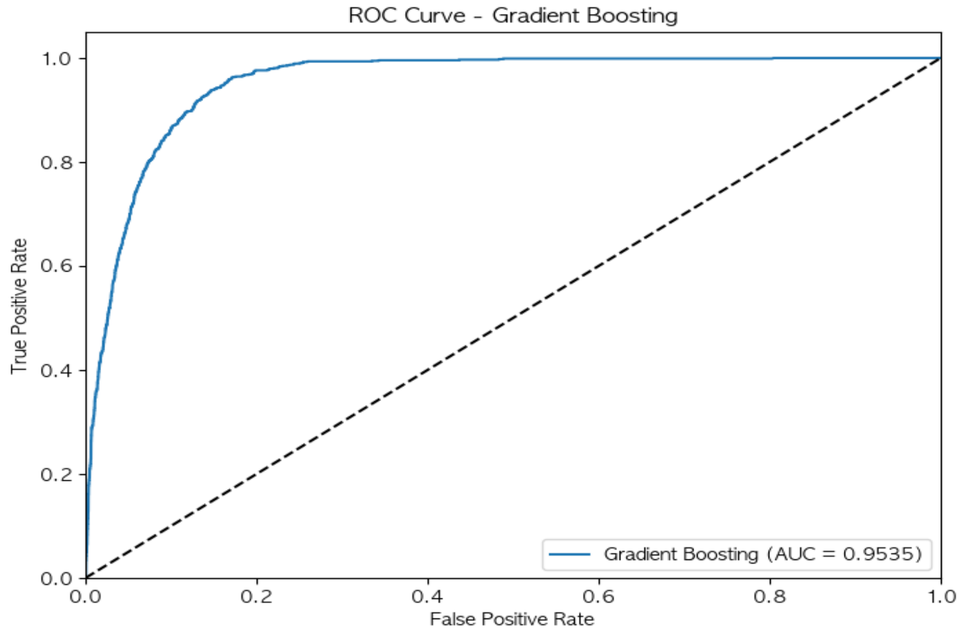
규칙 기반 모델 / 직관적인 해석

05 모델 성능 평가 및 결과



Random Forest

앙상블 기법 / 안정적 성능

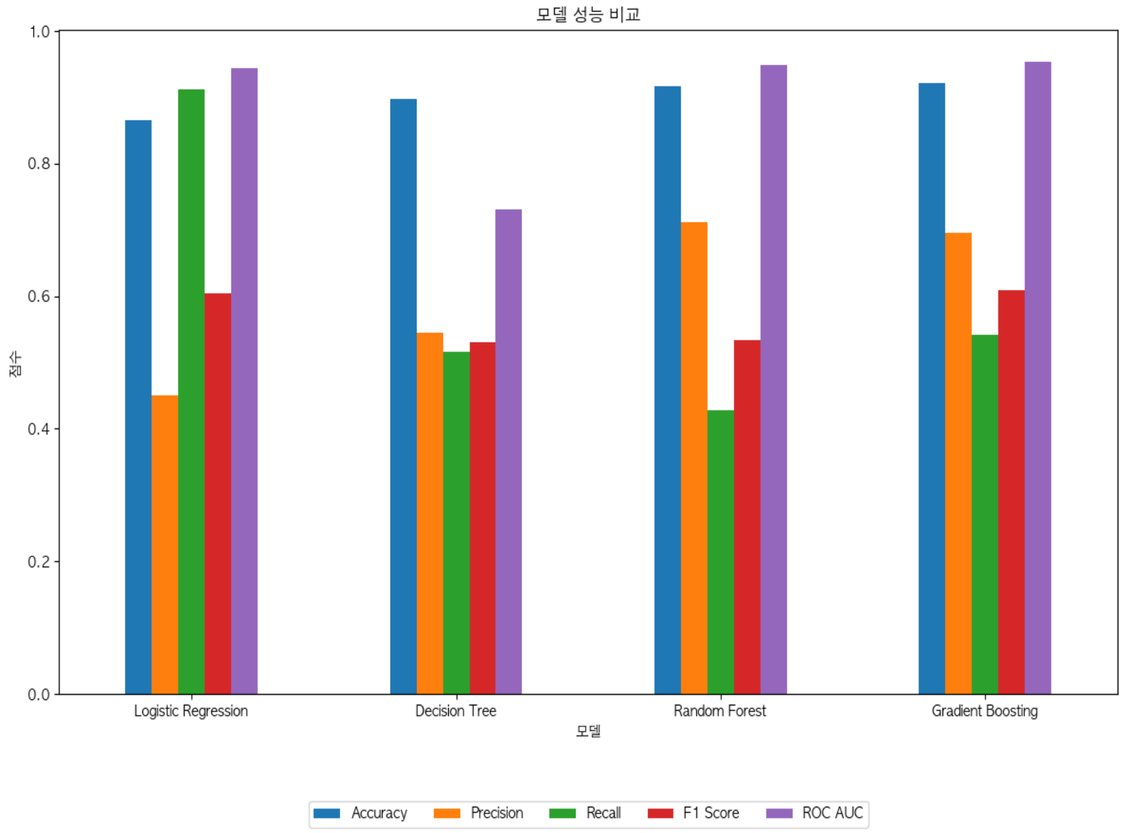


Gradient Boosting

순차적 학습 / 복잡한 패턴 학습

05 모델 성능 평가 및 결과

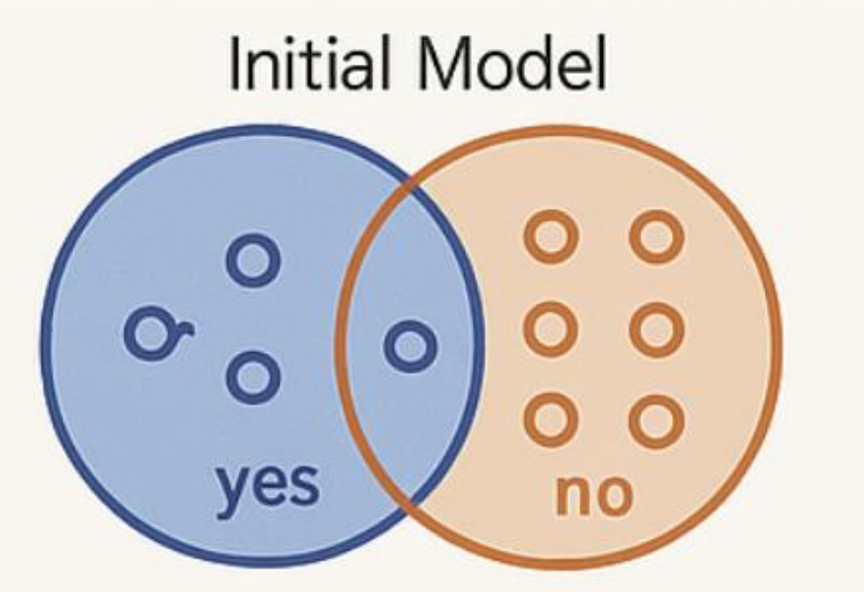
모델 성능 비교



모델 성능 비교:

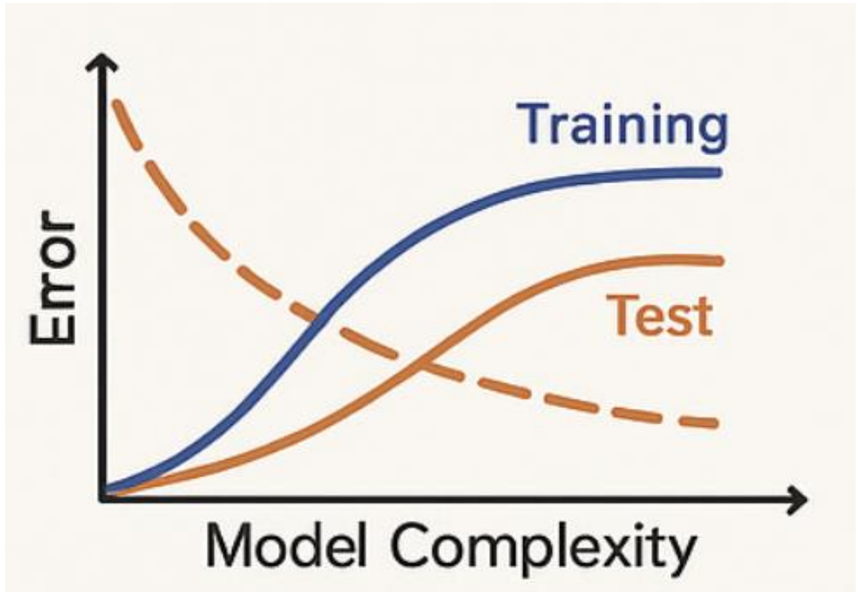
	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.865137	0.451200	0.911638	0.603639	0.943838
Decision Tree	0.897427	0.547106	0.519397	0.532891	0.732407
Random Forest	0.916485	0.715827	0.428879	0.536388	0.947866
Gradient Boosting	0.921704	0.695712	0.542026	0.609328	0.953484

06 문제점 및 해결 방안



클래스 불균형으로 인한 편향

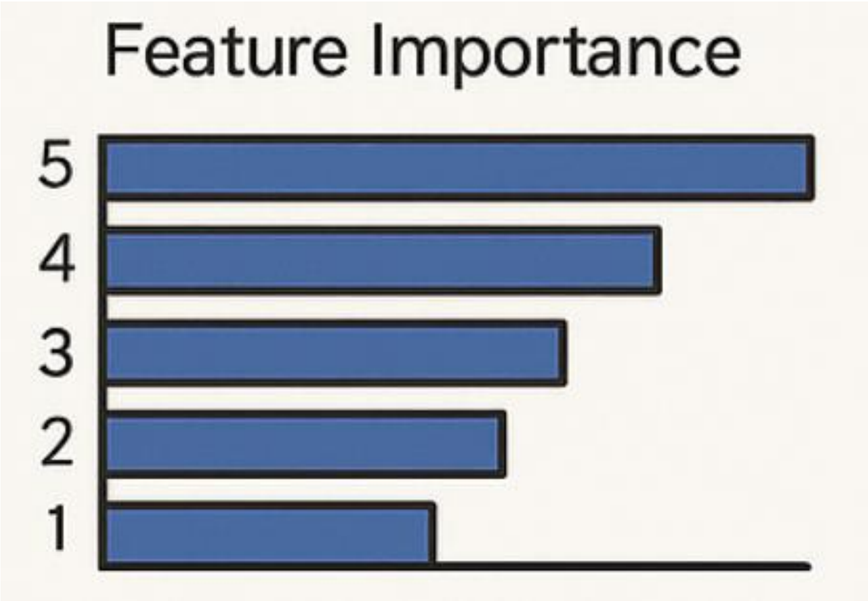
발견: 초기 모델이 다수 클래스(미가입)에만 편향
해결: class_weight='balanced' 파라미터 적용
효과: Recall 성능 대폭 향상



과적합 위험성

발견: 복잡한 모델에서 훈련/테스트 성능 차이
해결: 교차검증 및 정규화 적용
효과: 일반화 성능 안정화

06 문제점 및 해결 방안



특성 중요도 해석의 어려움

발견: 원-핫 인코딩 후 특성 수 급증 (62개)
해결: 상위 중요 특성 위주 분석
효과: 비즈니스 인사이트 도출 가능

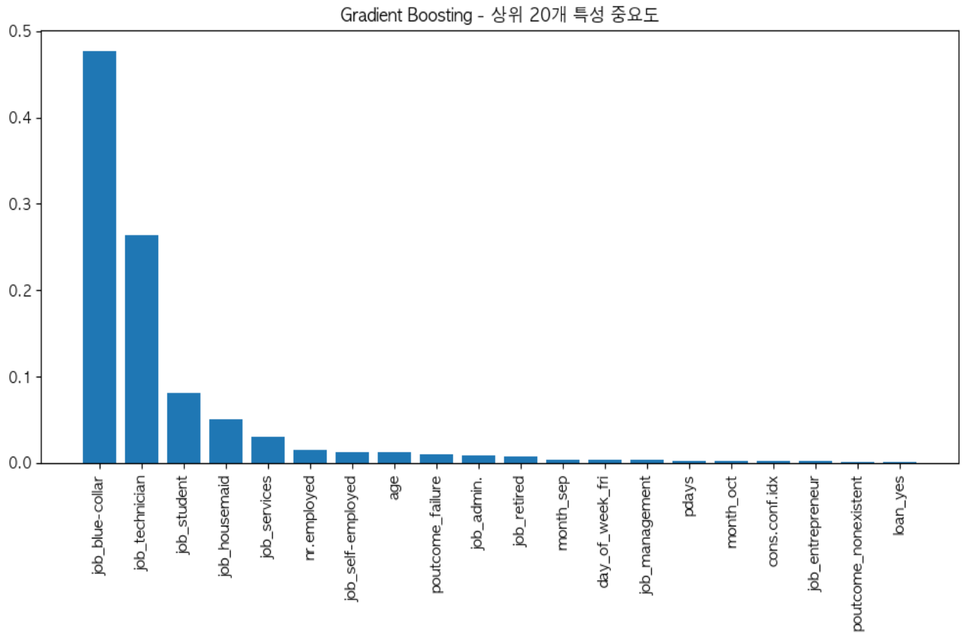
계산 시간 문제

발견: GridSearchCV로 인한 긴 실행 시간
해결: 파라미터 범위 축소, 병렬 처리 활용
효과: 효율적인 모델 최적화

07 결과 분석 및 논의

최고 성능 모델: Gradient Boosting

F1 Score : 0.6093(불균형 데이터 최적화)
정확도: 92.17%/ ROC-AUC: 0.9535
4개 모델 중 모든 지표에서 최고



주요 특성 변수

1. 직업이 육체 노동인 경우(job_blue-collar): 40%이상 기여 (0.48)
2. 직업이 기술직인 경우(job_technician): 높은 예측력 (약 0.26)
3. 직업이 학생인 경우 (job_student) : 약 0.08
4. 직업이 가사도우미인 경우 (job_housemaid): 약 0.05

비즈니스 인사이트: 직업군 기반 전략

최우선 타겟: Blue-Collar 직군

- 모델 기여도 최상위(약 0.48)
- 전략1: 육체 노동 직군 고객을 대상으로 특별 금리나 우대 혜택 제공
- 전략2: 간소화된 절차 + 짧은 상담 프로세스
  -> 고객이 쉽게 가입 유도

2차 타겟: Technician 직군

- 기술직 고객이 선호하는 워크플로우에 맞춘 앱이나 스크립트 제공
- 비용 대비 효율적 집중 관리

보조 타겟: 학생, 가사도우미 등 나머지 직업군

- 학생에게는 소액 정기예금 옵션 제공
- 가사도우미에게는 맞춤형 입금 주기 안내
- 해당 그룹은 별도의 세그먼트로 묶어 별도 프로모션 제공 고려

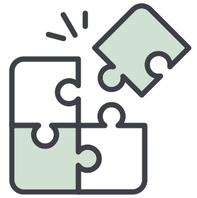
ROI 기대효과

비용 기대 수익률 개선	예측 모델 도입으로 불편한 마케팅 비용을 줄이고 높은 전환율을 달성하여 투자 대비 수익률이 크게 상승하는 데 기여할 수 있음
가입당 비용(CAC) 획기적 절감	고객 예측 확률 기준으로 고확률 그룹만 집중 공략하여 한명을 확보하는데 드는 비용을 크게 감소시킬 수 있음
순수익 증가	비용 절감뿐 아니라 예측 모델로 전환율이 향상되어 전체 순수익이 크게 상승할 수 있음
운영 효율성 강화	예측 결과 기반으로 고객을 세분화하여 상담원과 마케팅 채널 자원을 최적 배치가 가능함

07 한계점 및 개선 방향

한계점

클래스 불균형
전체 데이터 학습 : train_test_split 한 번만 사용
단일 평가 지표 : F1 Score 위주 최적화



개선 방향

- **K-fold 확장** : 5-fold or 10-fold로 안정성 향상
- **특성 조합** : 파생 변수 생성
- **다중 지표 최적화** : Precision-Recall 균형 고려
- **앙상블 모델** : 여러 모델 조합으로 성능 향상

감사합니다