

은행 정기 예금 가입

마케팅 전략 제안

예측 모델 및 고객 클러스터링 기반 인사이트 도출

분석 코드는 [여기\(클릭\)](#)서 확인하실 수 있습니다.

문상혁

INDEX

01 배경과 목적

02 EDA

기술통계분석
데이터 전처리
상관관계 분석
카테고리별 분석

03 Feature Engineering

파생변수 생성 및 선별
오버샘플링

04 예측 모델 개발

Random Forest
XGboost
LightGBM

05 클러스터링 분석 및 전략 제안

경제상황에 따른 클러스터별 인사이트

06 결론



배경

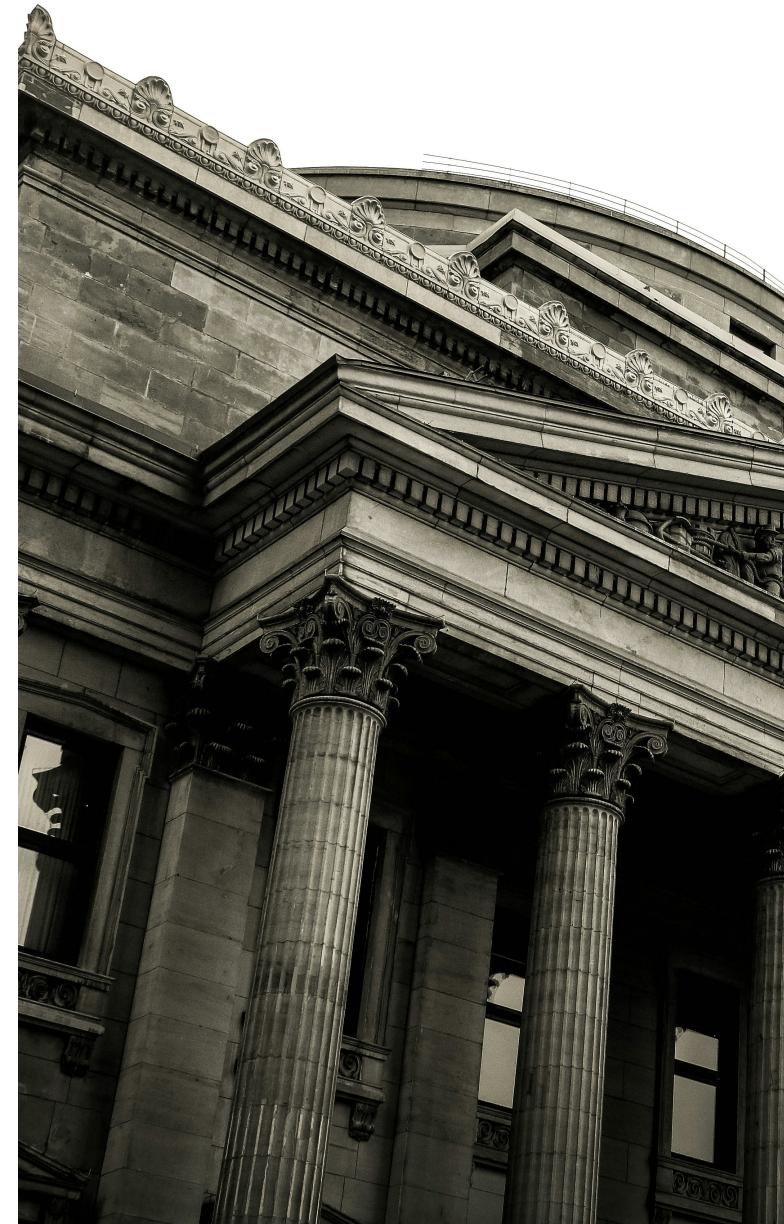
포르투갈 은행의 마케터로, 데이터를 기반으로 정기 예금 가입 가능성이 높은 고객을 식별하고, 고객 맞춤 마케팅 전략을 수립하여 불필요한 비용을 감소시키고 마케팅 효율성을 제고 해야 한다.

목적

결정 트리 및 양상블 기법을 활용하여 고객이 정기 예금에 가입할 가능성을 예측하는 분류 모델을 구축하여 마케팅 캠페인의 성공률을 높이고, 은행의 고객 확보 전략을 최적화하기 위함

목표

1. 결정 트리 및 랜덤 포레스트, 그래디언트 부스팅 등 다양한 양상블 모델을 활용하여 예측 성능을 비교하고 최적의 모델을 선택한다.
2. 예측 모델 기반으로 중요한 변수 선별하여 해당 변수 기반 클러스터링으로 고객 세그먼트 맞춤별 마케팅 전략을 도출한다.



- ✓ 2008년부터 2010년까지의 은행 마케팅 캠페인 데이터
- ✓ 종속변수인 정기 예금 가입 여부 컬럼까지 총 21개의 컬럼, 41188개 데이터 존재
- ✓ 회원 정보, 마케팅 캠페인 데이터, 경제 지표로 구분하여 탐색 진행

회원 정보

컬럼명	설명	비고
age	나이	int
job	직업	object
marital	결혼 여부	object
education	교육 수준	object
default	신용 불량 여부	object
housing	주택 대출 여부	object
loan	개인 대출 여부	object

마케팅 캠페인 데이터

컬럼명	설명	비고
contact	연락 유형	object
month	마지막 연락 월	object
day_of_week	마지막 연락曜일	object
duration	마지막 연락 지속 시간	int
campaign	캠페인 동안 연락 횟수	int
pdays	이전 캠페인 후 지난 일수	int
previous	이전 캠페인 동안 연락 횟수	int
poutcome	이전 캠페인 결과	object

경제 지표 및 종속변수

컬럼명	설명	비고
emp.var.rate	고용 변동률	int
cons.price.idx	소비자 물가 지수	object
cons.conf.idx	소비자 신뢰 지수	object
euribor3m	3개월 유리보 금리	object
nr.employed	고용자 수	object
y	정기 예금 가입 여부	object

표 1,2, 3. 데이터 컬럼별 설명

기술통계분석 – 데이터 분포 확인

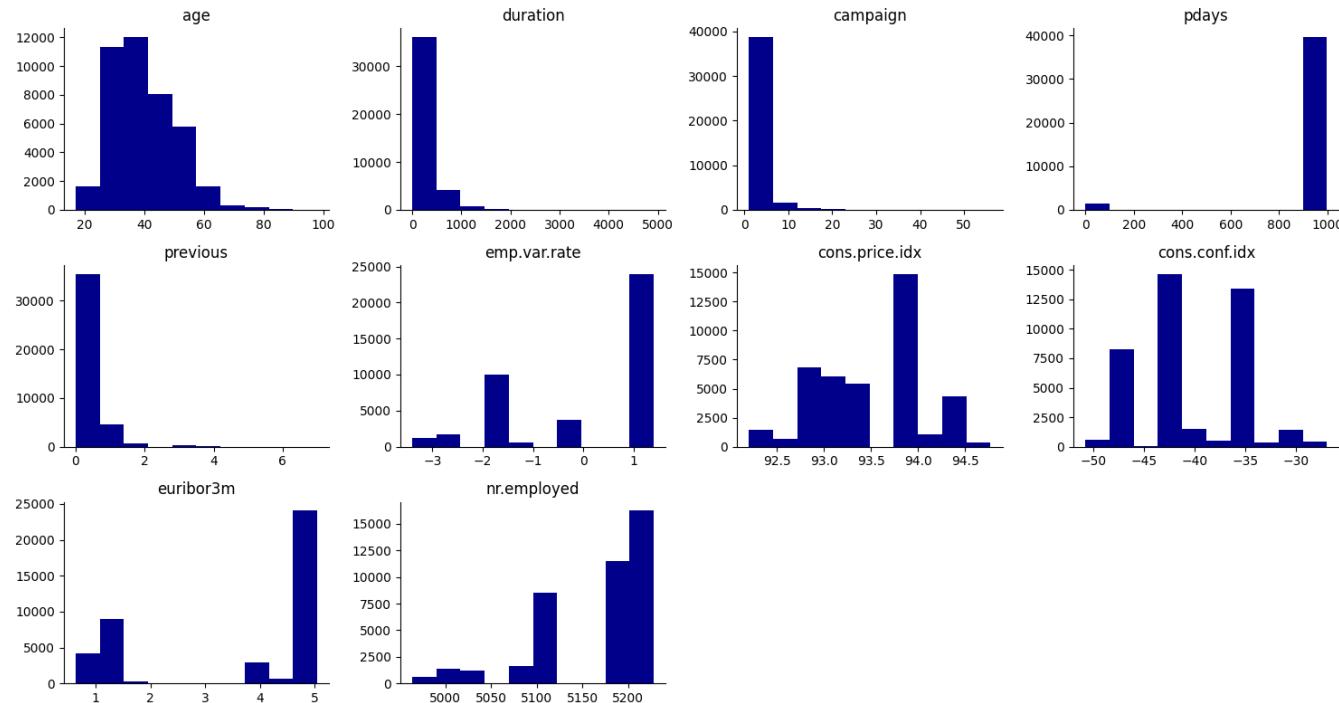
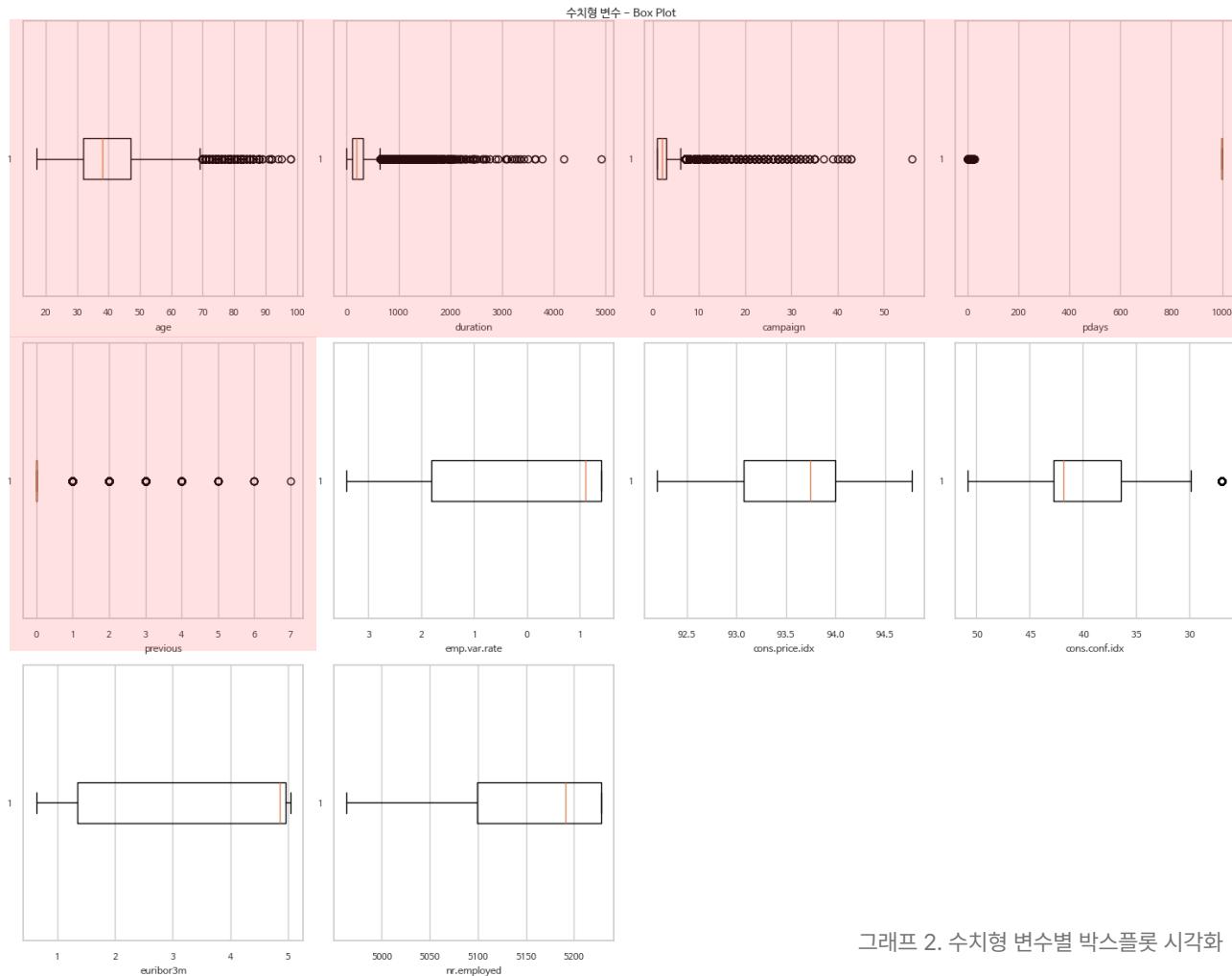


그림 1. 수치형 변수 데이터 컬럼별 분포 시각화

- **Pdays(이전 캠페인 이후 지난 일 수)**가 대부분 높은 수치로 편향되어 분포한다.
 - Pdays가 9990이면 한번도 연락하지 않은 고객임을 의미하기 때문에 데이터의 편향이 발생한 것임을 확인
- **3050 회원들이 다수를 차지하고 있으며, 20대 및 60대 이후 회원은 적게 분포한다.**
 - 경제 활동을 활발하게 하는 사람들이 은행에 자주 방문하기 때문이라고 추정 가능함

이상치 처리 – KNN 알고리즘을 통해 최근접 이웃들의 평균으로 이상치 대치



✓ 마케팅 캠페인 관련 데이터 이상치 다수 발생

구체적인 이상치 비율을 확인해보니 previous 컬럼에서 가장 많은 이상치 발생(13.66%)

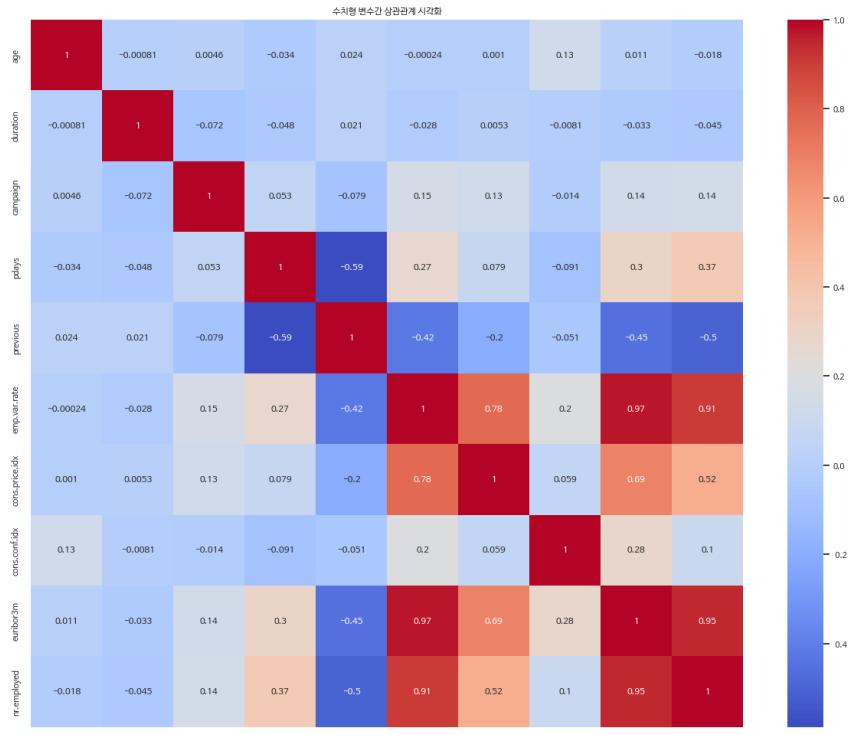
✓ 경제 지표 데이터는 이상치 거의 발생 X

경제 지표는 직접 데이터를 측정한 것이 아닌
통계청에서 발표한 데이터를 수집한 것일 가능성이 큼

✓ KNN 알고리즘 활용하여 이상치 처리

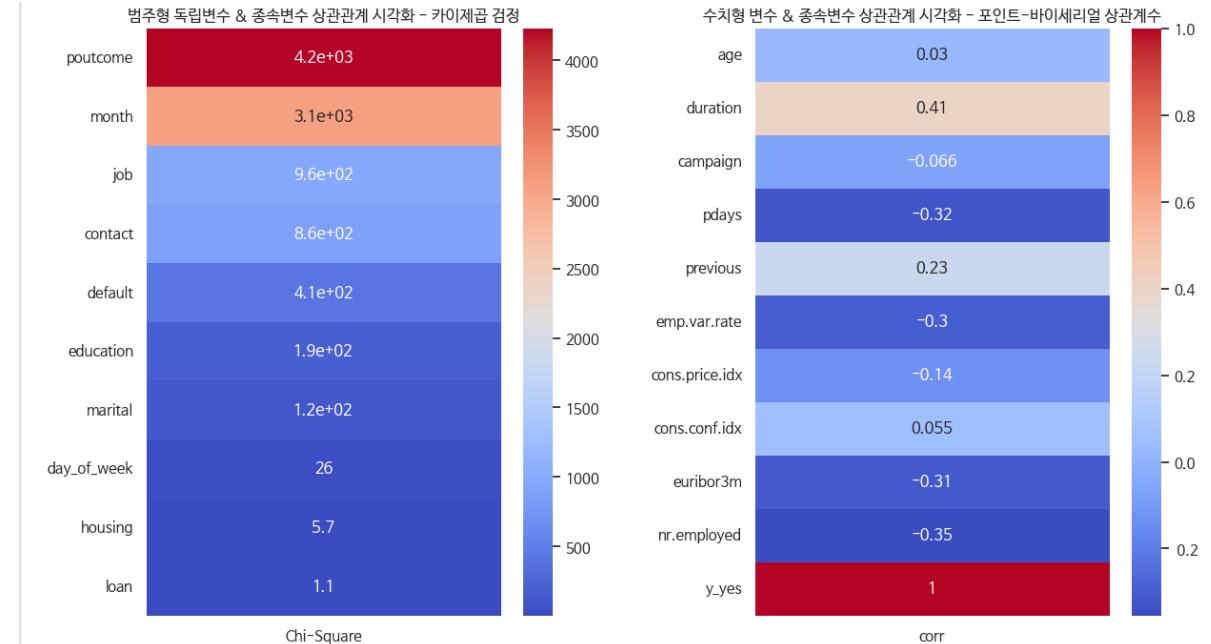
1. 사분위수를 활용한 이상치 탐지 결과 다수 탐지
2. 이상치에 강한 랜덤포레스트, 부스팅 모델 활용 예정
3. 예측 모델 개발을 위해서 데이터 최대한 보존 필요
4. 따라서 **KNN 알고리즘을 통해 모든 컬럼의 이상치를 처리**
(최근접 이웃들의 평균으로 이상치 대치)

상관관계 분석



그래프 3. 변수별 상관관계 시각화

- ✓ 변수별 상관관계 시각화 결과 독립변수 중 수치형 변수간 높은 상관관계 존재 확인
- ✓ 종속변수가 범주형 변수 : 독립변수 유형에 따라 계수 선정(피어슨 상관계수 X)
 - 수치형 변수 : 포인트 – 파이세리얼 상관계수
 - 범주형 변수 : 카이제곱 통계량



그래프 4. 종속변수와의 상관관계 시각화

✓ 상관관계 분석 결과(수치형 변수 – 포인트-파이세리얼 상관계수)

- 양의 상관관계(0.3 이상) : duration, previous
- 음의 상관관계(0.3 이상) : pdays, emp.var.rate, euribor3m, nr.employed

✓ 상관관계 분석 결과(범주형 변수 – 카이제곱 통계량)

- 가장 영향력 큰 변수(poutcome), 가장 영향력 작은 변수(day_of_week)
- 영향력이 큰 순서대로 시각화(빨강 → 파랑)
- P-value가 0.05보다 큰 변수 : housing, loan

카테고리별 변수 세부 분석 – 회원정보 / 마케팅 캠페인 / 경제 지표 데이터 별 종속변수와의 회귀분석으로 변수별 영향력을 살펴본다.

- ✓ 범주형 변수는 One-Hot Encoding을 통해 인코딩, 수치형 변수는 표준화
- ✓ VIF 계수 확인하여 10 (다중 공선성 존재) 이 넘는 변수와 inf(해당 변수가 완전히 다른 변수들의 선형 조합 = 다른 변수 조합으로 정확하게 표현 가능 = 다중 공선성 높음)인 변수 제거
- ✓ 종속변수 결과(예금 가입 O, X)에 따른 데이터 수 차이를 관찰하여 샘플링 방식 선정

회원 정보

- ✓ 연령, 직업, 결혼 여부 관련 변수들이 종속변수에 강한 영향
- ✓ **교육(education)**관련 변수는 대부분 **p-value**가 0.05보다 큼
- ✓ 모델 개발시 **교육(education)**은 추가 **Engineering** 필요 or 제거
- ✓ 유의미하지 않은 변수들은 (**p-value > 0.05**) 모델에서 제거(**housing, loan**)

마케팅 캠페인 데이터

- ✓ 통화한 시간(**duration**)이 길수록 정기예금 가입 여부에 **긍정적인** 영향
- ✓ 전화 연락(**contact_telephone**)은 정기예금 가입 여부에 **부정적인** 영향
- ✓ 월(**month**)마다 정기예금 가입 여부에 영향을 끼치는 정도가 다르다.
- ✓ 정기예금 가입 여부에 영향을 미치는 특정 **요일(tue, wed)** 존재, 나머지 요일은 **p-value > 0.05**

경제 지표 및 종속변수

경제 지표

- ✓ 소비자 물가 지수(**cons.price.idx**) 정기예금 가입 여부에 **긍정적인** 영향
- ✓ 소비자 신뢰 지수(**cons.price.idx**) 정기예금 가입 여부에 **부정적인** 영향

종속 변수 – **X** 데이터 오버 샘플링 필요

정기 예금 가입 여부	데이터 수
O	36,529
X	4,635

Feature Engineering 과정 요약

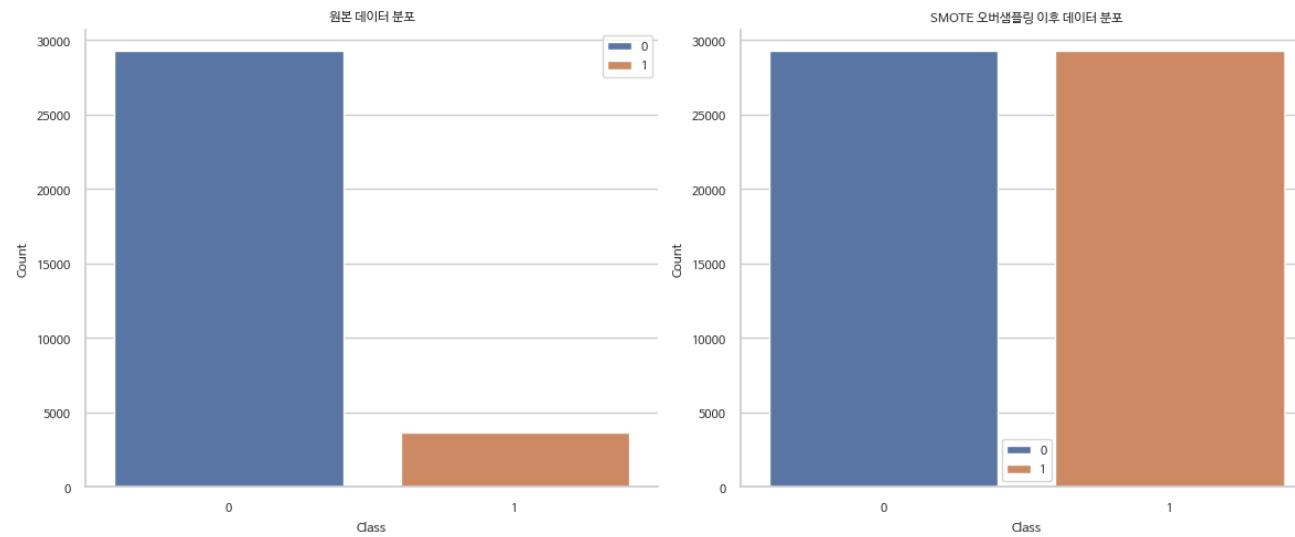
- ✓ 결측값과 중복값, 이상치는 EDA 과정에서 처리 완료
- ✓ pdays변수에서 파생변수 생성
 - pdays의 999값은 한번도 연락하지 않았다는 뜻이므로 데이터 왜곡 발생 가능성 존재
 - 따라서 pdays값을 일정한 범주를 나누어 처리
- ✓ 범주형 변수 One-Hot Encoding 진행
- ✓ 종속변수의 정기 예금 미가입 데이터 오버 샘플링(SMOTE 활용)

파생변수 생성 – pdays 범주화

pdays	설명
0~6	1
7~13	2
14~20	3
21~27	4
999	0

표 4. pdays 변수 범주화

SMOTE 기법으로 종속변수(정기 예금 미가입 데이터) 오버샘플링



Random Forest

- ✓ 그리드 서치 활용하여 최적 파라미터 탐색 - max_depth: None, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 100
- ✓ Accuracy, Precision, F1 Score, ROC AUC 지표를 통해 성능 측정 - 0.9547 / 0.9548 / 0.9554 / 0.9936
- ✓ 속성 중요도 TOP 10

duration, euribor3m, nr.employed, emp.var.rate, cons.conf.idx, cons.price.idx, age, contact_telephone, poutcome_nonexistent, job_blue-collar

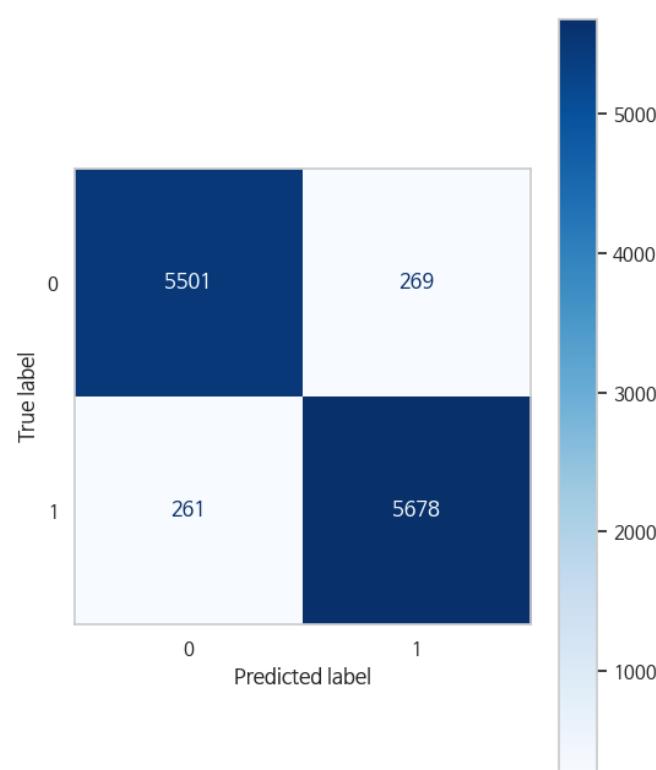


그림 6. Confusion Matrix of Random Forest model

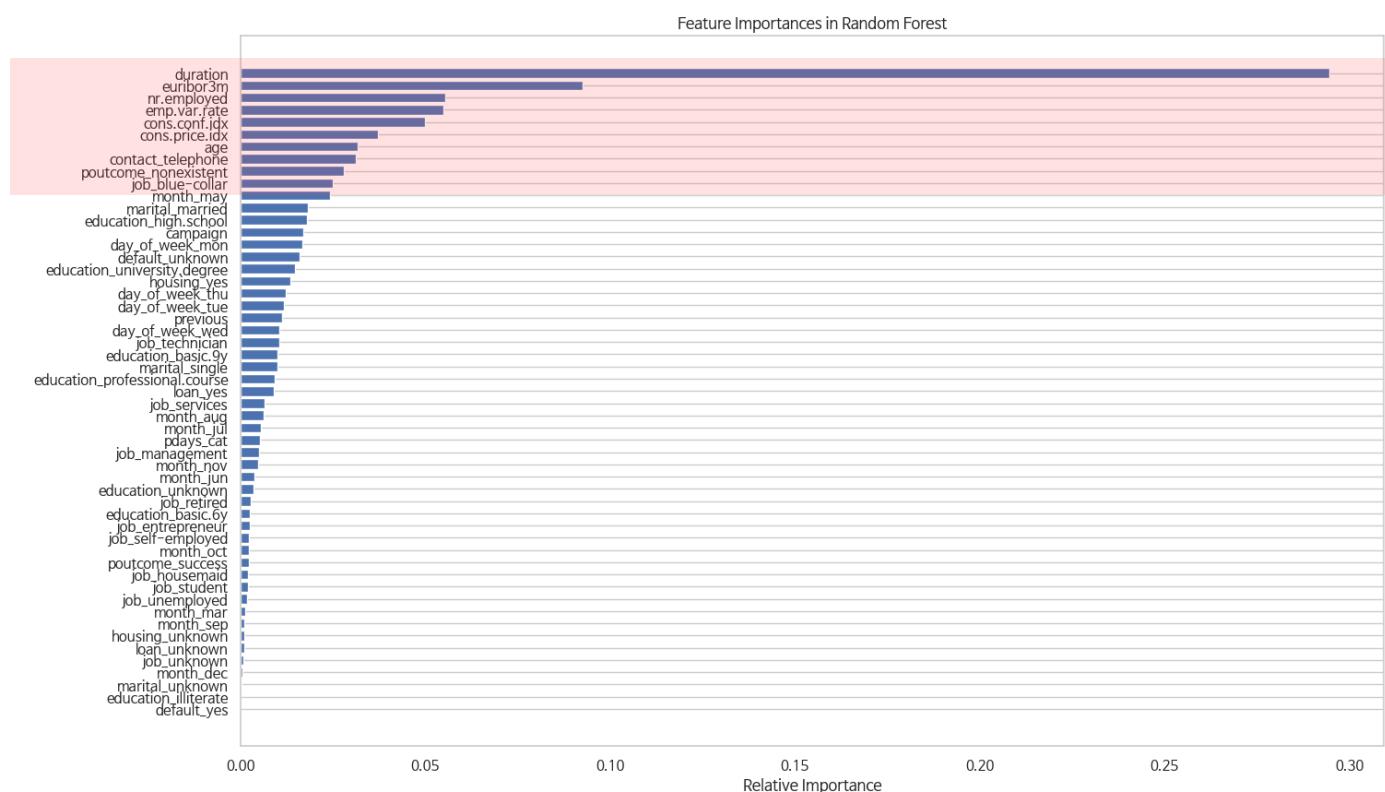
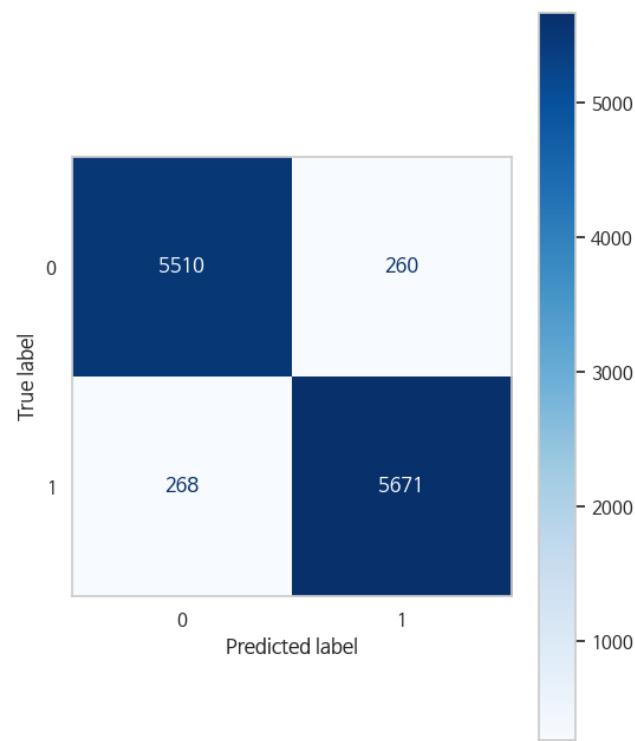


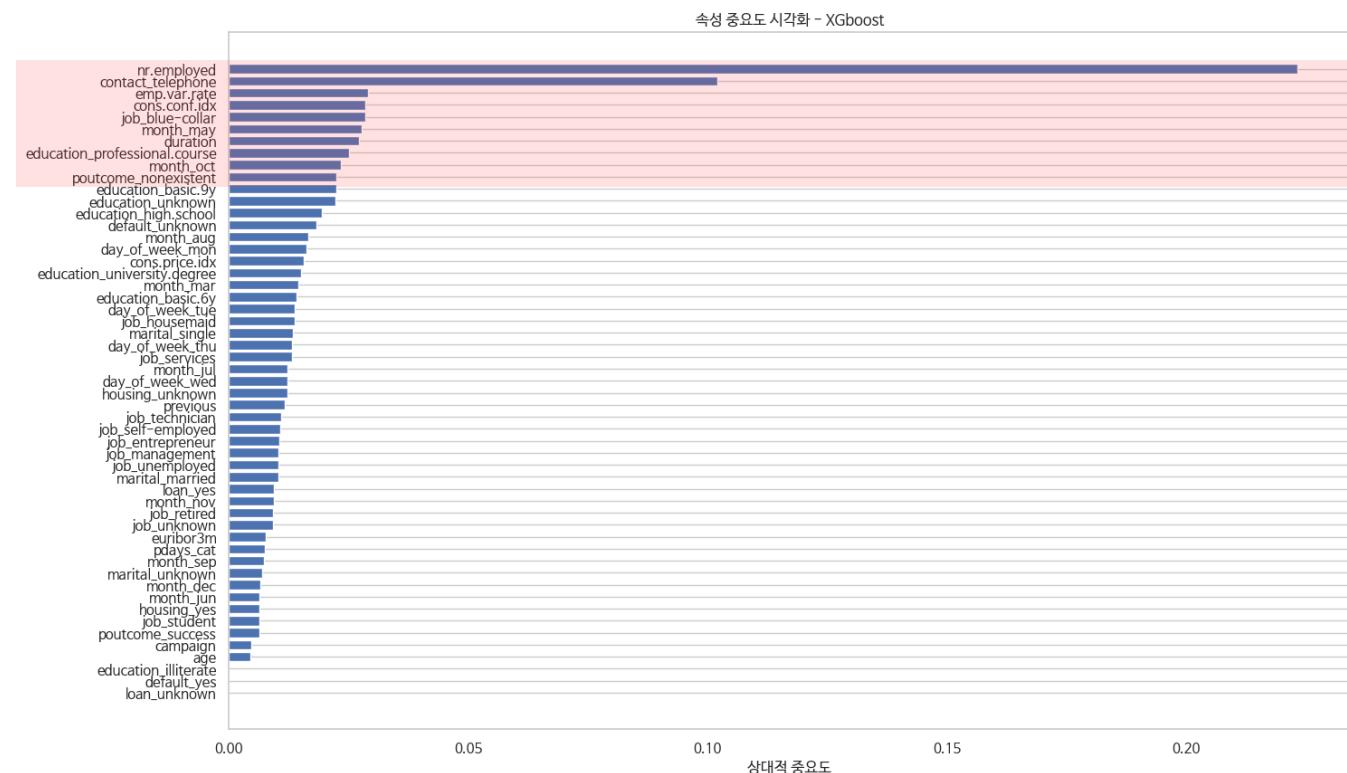
그림 7. Random Forest model 속성 중요도

XGboost

- ✓ 그리드 서치 활용하여 최적 파라미터 탐색- colsample_bytree: 1.0, learning_rate: 0.1, max_depth: 7, n_estimators: 500, subsample : 0.6
- ✓ Accuracy, Precision, F1 Score, ROC AUC 지표를 통해 성능 측정 – 0.9549 / 0.9562 / 0.9555 / 0.9933
- ✓ 속성 중요도 TOP 10 - nr.employed, contact_telephone, emp.var.rate, cons.conf.idx, job_blue-collar, month_may, duration, education_professional.course, month_oct, pcoutcome_nonexistent



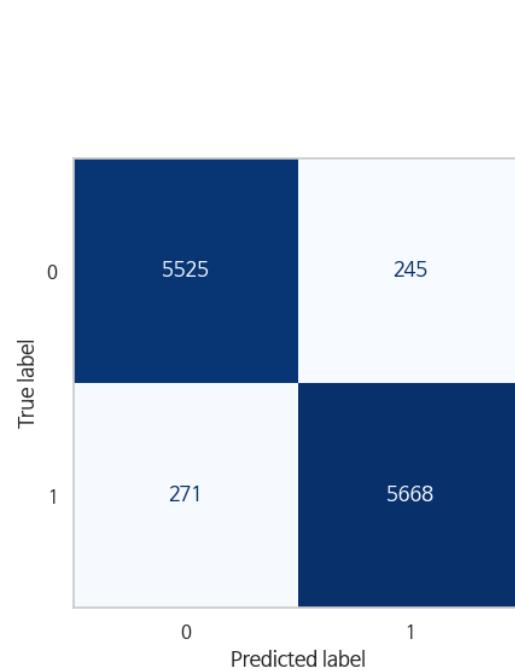
그리프 8. Confustion Matrix of XGboost model



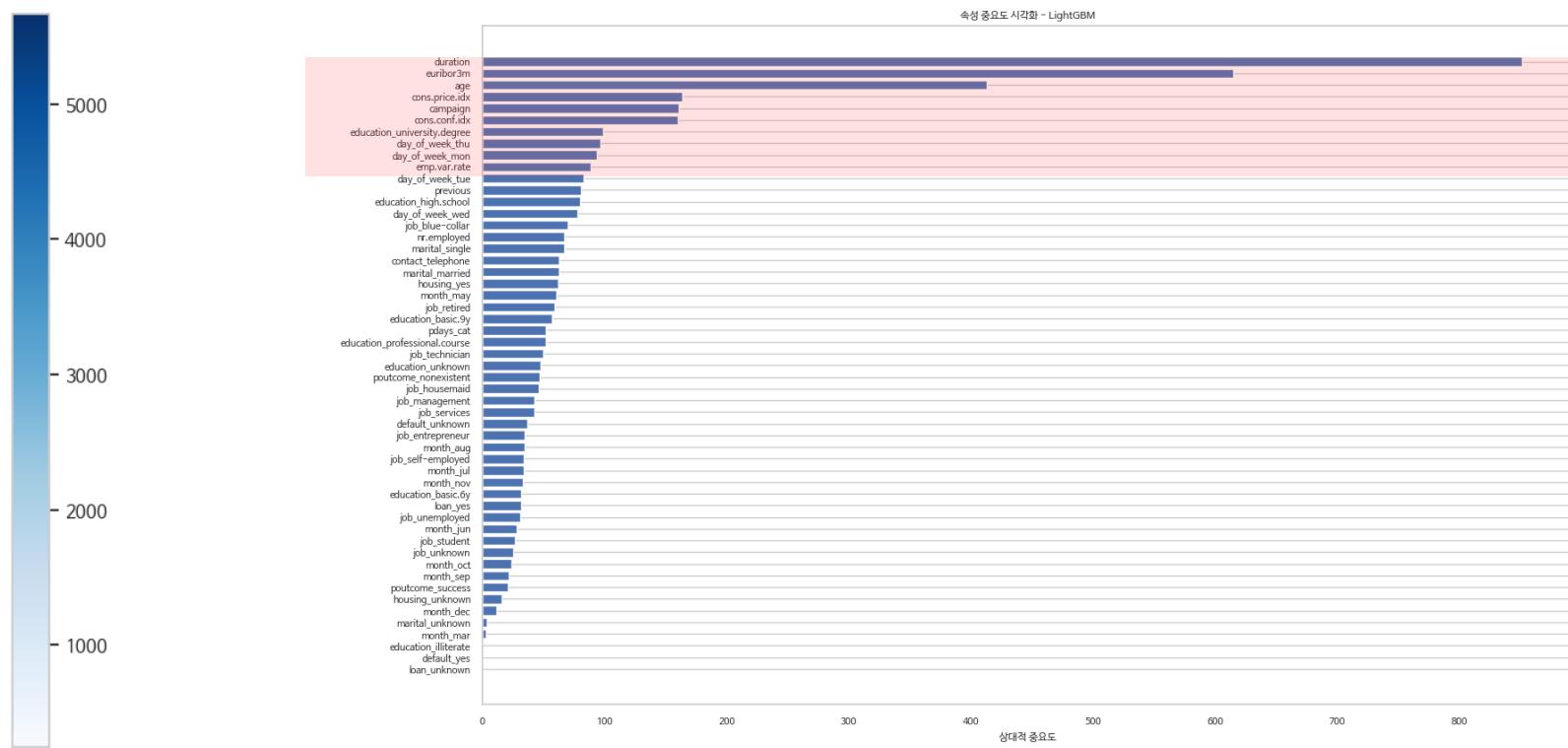
그리프 9. XGboost model 속성 중요도

LightGBM

- ✓ 그리드 서치 활용하여 최적 파라미터 탐색 - colsample_bytree: 1.0, learning_rate: 0.1, max_depth: -1, n_estimators: 150, num_leaves: 31, subsample: 0.6
 - ✓ Accuracy, Precision, F1 Score, ROC AUC 지표를 통해 성능 측정 – 0.9559 / 0.9586 / 0.9565 / 0.9934
 - ✓ 속성 중요도 TOP 10
- duration, euribor3m, age, cons.price.idx, campaign, cons.conf.idx, education_university.degree, day_of_week_thu, day_of_week_mon, emp.var.rate



그래프 10. Confusion Matrix of LightGBM model



그래프 11. LightGBM model 속성 중요도

- ✓ 각 모델의 속성 중요도 결과를 참고하여 클러스터링 수행할 속성 선별
- ✓ 각 모델 **TOP10 교집합**으로 속성 선별 - 각 모델 성능에 따라 가중치를 두고 진행하려 했으나, 모델의 성능이 거의 비슷하여 단순 교집합으로 선별
- ✓ 세 모델 모두 공통적으로 중요하게 사용된 속성 - **duration, emp.var.rate, cons.conf.idx**
 - 클러스터링을 수행하기엔 위 3개의 속성이 적다고 판단되어 두 모델에서 공통적으로 중요하게 사용된 속성 중 2개 추가 선별
- ✓ 두 모델 공통적으로 중요한 속성 중에 가장 중요도가 높은 속성 2개를 추출하기 위해 점수 시스템 도입
 - 각 모델에서 중요도 순대로 1~10점 스코어링
 - 모델간 성능이 비슷하므로 가중치를 두지 않음
 - 각 모델의 해당 특성의 점수 총합으로 TOP2 선별(ex. age는 LightGBM에서 3위로 8점, xgboost에서 7위로 4점 따라서 총점 12점)

두 모델에서 공통적으로 중요한 속성 스코어 보드

속성	점수
nr.employed	18
euribor3m	18
contact_telephone	12
cons.price.idx	12
age	12
job_blue-collar	7

선정된 속성

속성	분류
duration	마케팅 캠페인
emp.var.rate	경제 지표
cons.conf.idx	경제 지표
nr.employed	경제 지표
euribor3m	경제 지표

표 6. 점수표 기반 선정된 속성



선정된 속성 5개 중 4개가 경제 관련 지표

1. 경제 지표는 우리가 통제할 수 없는 지표
2. 해당 특성을 기준으로 클러스터링을 수행해서 고객을 군집화 하면, 경제 상황에 따른 전략 외 실효성있는 인사이트 도출에 무리가 있다고 판단
3. 경제 지표 제외하고 속성 재선정
 - **duration, age, contact_telephone, job_blue-collar**
4. 경제 지표를 활용하여 경제 상황이 좋은(A), 보통(B), 나쁨(C) 그룹으로 나누어
 - 경제 상황이 동일한 그룹 내에서 클러스터링 수행

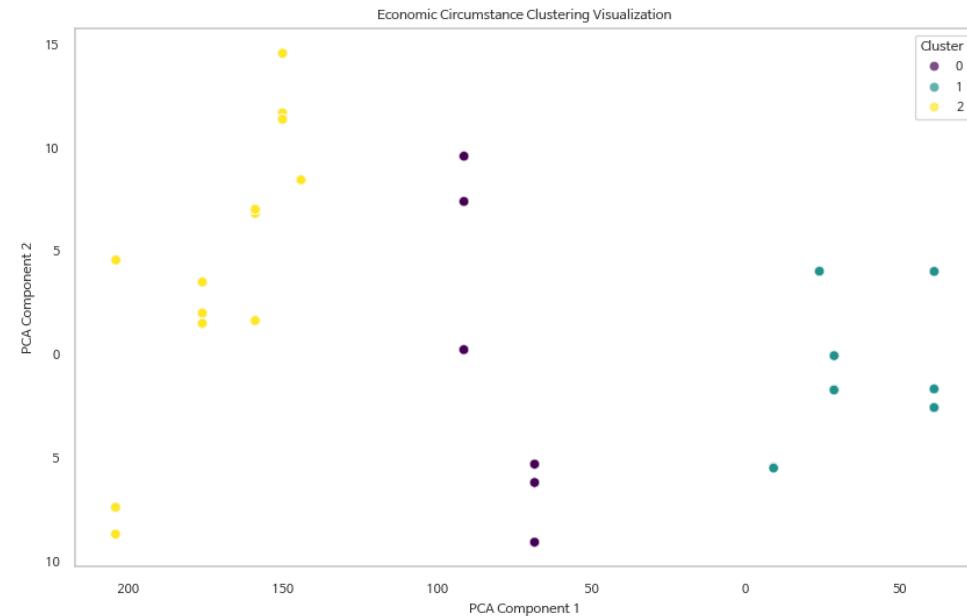
표 5. 두 머신러닝 모델에서 공통적으로 중요한 속성들의 점수표

K-MEANS 클러스터링으로 경제 상황별 그룹 분류

✓ 경제 지표 활용 - emp.var.rate / cons.price.idx / cons.conf.idx / euribor3m / nr.employed

속성	분류
그룹 0	경제 상황 보통
그룹 1	경제 상황 좋음
그룹 2	경제 상황 나쁨

표 7. 경제상황 클러스터링 결과



그래프 12. 경제 상황 클러스터링 시각화

- 그룹 0: emp.var.rate가 -1.98로 비교적 낮고, nr.employed도 중간 수준 → 보통 그룹
- 그룹 1: emp.var.rate가 가장 높고, nr.employed도 가장 많으며, euribor3m도 가장 높음 → 좋음 그룹
- 그룹 2: emp.var.rate가 가장 낮고, nr.employed도 가장 적음 → 나쁨 그룹

K-MEANS 클러스터링으로 고객군 분류

- ✓ 앞서 선정한 속성 활용 - **duration**, **age**, **contact_telephone**, **job_blue-collar**
- ✓ 경제 상황 별로 Inertia를 시각화하여 최적 클러스터 수 선정
- ✓ 경제 상황 별로 클러스터링 수행 후 클러스터 간 평균 차이가 두드러지는 특성 선별하여 해석 진행

클러스터	경제 상황 보통	경제 상황 높음	경제 상황 낮음
경제 활동의 주축	<ul style="list-style-type: none"> - 예금 가입률 50.8% (가장 높음) - 장시간 상담 가능 (818.5초) - 기존 캠페인 성공률 가장 높음 (6.43%) - 30대 중반 연령층 (37.25세) - 미혼 비율 36.49% 	<ul style="list-style-type: none"> - 예금 가입률 24.1% (중하) - 평균 연령 38.96세 (중간) - 전화 연락 비율 45.99% (다소 낮음) - 은퇴자 비율 2.05% (중간) - 미혼 비율 26.62% (중간) - 통화 시간 904.15초 (가장 길) 	<ul style="list-style-type: none"> - 예금 가입률 73.36% (가장 높음) - 평균 연령 41.38세 (중간 연령층) - 전화 연락 비율 13.82% (낮음) - 은퇴자 비율 9.54% (낮음) - 기혼 비율 57.57% (중간) - 대학 교육 비율 42.43% (중간)
사회 초년생 / 재테크 입문자	<ul style="list-style-type: none"> - 예금 가입률 11.3% (가장 낮음) - 평균 연령 32.89세 (젊은 층) - 전화 상담 시간 짧음 (189.74초) - 미혼 비율 가장 높음 (44.49%) - 전화 연락 비율 7.83%로 가장 높음 	<ul style="list-style-type: none"> - 예금 가입률 18.6% (가장 낮음) - 평균 연령 33.58세 (젊은 층) - 전화 연락 비율 50.47% (가장 높음) - 은퇴자 비율 0.14% (거의 없음) - 미혼 비율 35.75% (높음) - 통화 시간 187.12초 (가장 짧음) 	<ul style="list-style-type: none"> - 예금 가입률 45.03% (중하) - 평균 연령 32.78세 (젊은 층) - 전화 연락 비율 19.15% (상당히 높음) - 은퇴자 비율 0.1% (거의 없음) - 기혼 비율 39.03% (중간) - 대학 교육 비율 45.76% (상당히 높음)
중장년층의 은퇴 대비자 / 고령의 노후 대비자	<ul style="list-style-type: none"> - 예금 가입률 15.1% (중간 수준) - 평균 연령 52.4세 (가장 높음) - 은퇴자 비율 가장 높음 (14.67%) - 소비자 신뢰도 중간 수준 - 통화 시간 208.32초 (중간) 	<ul style="list-style-type: none"> - 예금 가입률 36.5% (중간) - 평균 연령 49.64세 (중장년층) - 전화 연락 비율 49.05% (높음) - 은퇴자 비율 6.84% (중간) - 미혼 비율 8.71% (낮음) - 통화 시간 192.35초 (짧음) 	<ul style="list-style-type: none"> - 예금 가입률 49.06% (중간) - 평균 연령 63.21세 (고연령층) - 전화 연락 비율 12.41% (낮음) - 은퇴자 비율 43.99% (상당히 높음) - 기혼 비율 75.56% (높음) - 대학 교육 비율 23.97% (낮음)

표 8. 고객 클러스터링 결과 해석 표

고객군별 마케팅 인사이트 도출 – 경제 상황 그룹



경제 활동의
주축

- AM의 개별 관리 전략: 상담에 충분히 시간을 투자하는 특성이 있으므로, 담당 Account Manager이 집중 관리하며 팔로우업
- 리타겟팅(리마케팅): 과거 가입 고객에게 이메일 및 카카오톡 활용하여 맞춤 혜택 홍보
- 디지털 마케팅 활용: 이메일/앱 푸시 알림을 통한 프로모션 진행

30대 중반의 젊은 층

자동 재예치 옵션 제공

고이율 정기 예금



사회초년생
재테크 입문자

- 단기 혜택 강조: 예금 이자율 인상, 가입 시 캐시백 프로모션
- 소액 상품 추천: 초기 부담이 적은 소액 정기 예금 상품 홍보
- 디지털 마케팅 집중: 전화보다 이메일, SNS, 앱 푸시 활용

2030 사회 초년생

자동이체 시 보너스 금리 적용

단기(1년 미만) 소액 정기 예금



은퇴/노후
대비자

- 장기 예금 상품 강조: 안정적인 금융 상품 선호, 노후준비 소구점 활용
- 세금 혜택 및 연금 연계 상품 홍보: 장기 가입 유도
- 오프라인 상담 유도: 은행 방문 상담 혜택 제공

평균 연령 52.4세

장기(3년 이상) 고이율 정기 예금

증여 우대형 상품

연금 연계 예금 상품

고객군별 마케팅 인사이트 도출 – 경제 상황 **좋음** 그룹



경제 활동의
주축

- 긴 상담 활용: 경제 성장과 수익 증대 가능성을 강조하며, 장기적인 투자 혜택 설명
- 신용 정보 제공 및 관리: 신용 관리와 함께 신뢰를 바탕으로 장기 투자 유도
- 장기 성장 상품 추천: 주식형 예금 등 성장 가능성을 고려한 투자 상품 제안

30대 중반의 경제 주축

장기(3년 이상) 고수익 정기 예금

신용 관리와 투자 교육 패키지



사회초년생
재테크 입문자

- 단기 고수익 혜택 제공: 예금 이자율을 높이고, 가입 시 즉시 보너스 제공
- 소액 정기예금 상품 홍보: 낮은 금액으로 시작할 수 있는 상품 홍보
- 디지털 마케팅 집중: 짧은 층을 겨냥해 SNS, 앱, 이메일로 홍보 채널 선정

2030 사회 초년생

단기(1년 미만) 고수익 정기 예금

짧은 층 맞춤 소액 투자 상품



은퇴/노후
대비자

- 전화 상담 활용: 전화 응답률이 높아 상담을 통해 즉각적인 혜택 제공으로 후킹
- 안정성과 고수익성 강조: 긍정적인 경제 상황을 반영한 고이율 상품 홍보
- 은퇴 후 자산 관리 플랜 컨설팅 서비스 제공: 은퇴를 대비하여 어떻게 자산을 관리해야 하는지 전문 컨설팅 제공

평균 연령 49.6세

중기(1~3년) 고이율 정기 예금

투자형 예금 상품

고객군별 마케팅 인사이트 도출 – 경제 상황 나쁨 그룹



경제 활동의
주축

- 금리 인하 대비 고정금리 상품 강조: 경제 불황에 대비하여 고정금리 상품으로 장기적인 안전성 강조
- 조기 해지 시 불이익 최소화: 유연한 해지 옵션 제공, 유동성 문제 해결을 위한 상품 설계
- 안정성과 보장성을 강조: 원금 보장이 강한 상품 홍보

30대 중반의 젊은 층

고정금리 정기 예금

원금 보장형 금융 상품

보험 연계 상품



사회초년생
재테크 입문자

- 저위험/고수익 상품 강조: 금리가 낮은 상황에서, 최소한의 리스크로 수익을 얻을 수 있는 상품 제공
- 소액 투자 상품: 자금 여유가 적은 고객을 겨냥해 적은 금액으로 시작할 수 있는 상품 추천 / 분산 투자 옵션 제공: 예금 외에도 다양한 소액 투자 상품 제공
- 원금이 보장되고, 예금자보호법에 보호받는 정기예금의 핵심 소구점인 안정성 강조하여 입문자에게 적합함을 강조

2030 사회 초년생

금리 우대형 상품

분산 투자 상품

소액 정기 예금



은퇴/노후
대비자

- 안정적인 투자 강조: 경제 불확실성에 대비하여 안정적인 고정금리 상품 홍보
- 연금 연계 상품 강화: 장기적으로 안정적인 수익을 원하는 은퇴자를 겨냥한 연금 상품 추천
- 디지털 채널 강화: 은퇴자가 온라인 서비스를 활용하기 어려울 수 있으므로, 고객 맞춤형 지원 채널 제공

평균 연령 63.2세

고정금리 정기 예금

연금 연계 상품

재테크 지원 서비스

1) 제언 및 시사점

- 모델 성능 측면: LightGBM이 가장 높은 예측 성능을 보였으며, 주요 영향 변수로 duration, contact_telephone, age, job_blue-collar 등이 확인되었다.
 - 향후 유사한 분석에서는 해당 변수를 중심으로 추가적인 인사이트를 도출하는 것이 효과적일 것으로 판단된다.
- 클러스터링 기반 마케팅 전략
 - 은행이 통제할 수 없는 경제 상황이라는 변수의 영향을 줄이기 위해 경제 상황별 고객군을 나누어, 분석의 정확도를 높였다.
- 실무 적용 가능성: 본 연구 결과를 실제 은행 마케팅 캠페인에 적용할 경우, 기존의 일괄적인 마케팅 방식에서 벗어나 보다 정밀한 타겟팅이 가능해질 것으로 기대된다.

2) 분석의 한계점

- 데이터 편향: pdays 변수의 극단값(999)의 영향으로 인해 일부 데이터가 편향될 가능성이 존재하며, 이를 보정하는 방식에 따라 모델의 성능이 달라질 수 있다.
- 고객 행동 데이터 부족: 마케팅 캠페인 기록과 기본 고객 정보를 활용하였으나, 고객의 은행 거래 내역(예: 입출금 패턴, 대출 기록 등)이 포함되지 않아 분석의 정밀도가 다소 제한적이었다.

3) 회고 및 향후 발전 방향

- 고객 행동 데이터 추가 활용: 향후 분석에서는 실제 은행 거래 데이터와 연계하여 보다 정밀한 고객 분석이 가능하도록 확장할 필요가 있다.
- 딥러닝 기반 모델 적용: 랜덤 포레스트, XGBoost, LightGBM 등의 머신러닝 모델을 활용하였으나, 딥러닝 기반의 모델(LSTM, Transformer 등)을 적용해 분석을 시도할 수 있다.
- 실제 마케팅 캠페인 테스트: 제안된 마케팅 전략을 A/B 테스트를 통해 검증하고, 실제 캠페인 성과를 비교 분석하여 전략의 효과성을 평가하는 것이 중요하다.

END

감사합니다.

문상혁