

# 스프린트 미션 15

💡 코드잇 데이터분석 4기 고명진

## 고객의 은행 예금상품 가입 전환 마케팅 제안

### 1. 분석 배경

포르투갈 은행의 마케팅 데이터를 분석하면서 동시에 은행에서 진행하는 예금상품 가입 캠페인을 통해서 예금을 가입할 고객을 분류하는 모델 생성 및 이에 따른 마케팅 방안을 제안합니다.

### 2. 분석 주제

제공받은 은행 데이터셋을 통해 고객이 정기 예금을 가입할 가능성을 예측하고, 이를 통해 마케팅 캠페인의 효율성을 높이는 것

분석을 통한 최종 목표는 가장 정확한 분류 모델을 개발하여 고객이 정기 예금을 가입할지 여부를 예측하고, 모델을 통한 비즈니스 인사이트를 제시

### 3. 데이터 설명

해당 데이터는 UC Irvine Machine Learning Repository에서 제공하는 **Bank Marketing(링크)** 데이터 입니다.

#### 3-1. 데이터 출처

Moro, S., Cortez, P., & Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, <http://dx.doi.org/10.1016/j.dss.2014.03.001>

#### 3-2. 데이터 컬럼 설명

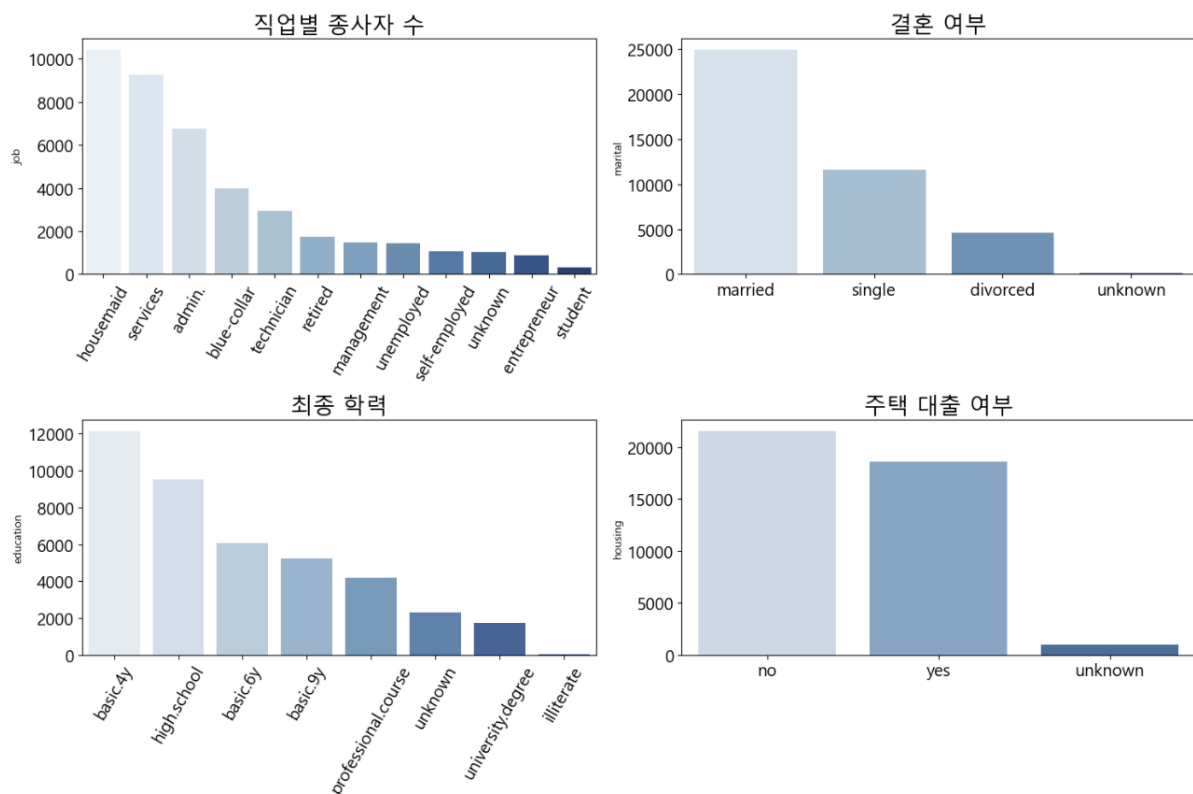
컬럼명	설명	컬럼명	설명
age	나이 (숫자)	duration	마지막 연락 지속 시간, 초 단위 (숫자)
job	직업 (범주형)	campaign	캠페인 동안 연락 횟수 (숫자)
marital	결혼 여부 (범주형)	pdays	이전 캠페인 후 지난 일수 (숫자)
education	교육 수준 (범주형)	previous	이전 캠페인 동안 연락 횟수 (숫자)

컬럼명	설명	컬럼명	설명
default	신용 불량 여부 (범주형)	poutcome	이전 캠페인의 결과 (범주형)
housing	주택 대출 여부 (범주형)	emp.var.rate	고용 변동률 (숫자)
loan	개인 대출 여부 (범주형)	cons.price.idx	소비자 물가지수 (숫자)
contact	연락 유형 (범주형)	cons.conf.idx	소비자 신뢰지수 (숫자)
month	마지막 연락 월 (범주형)	euribor3m	3개월 유리보 금리 (숫자)
day_of_week	마지막 연락 요일 (범주형)	nr.employed	고용자 수 (숫자)
		y	정기 예금 가입 여부 (이진: yes=1, no=0)

## 4. 탐색적 데이터 분석

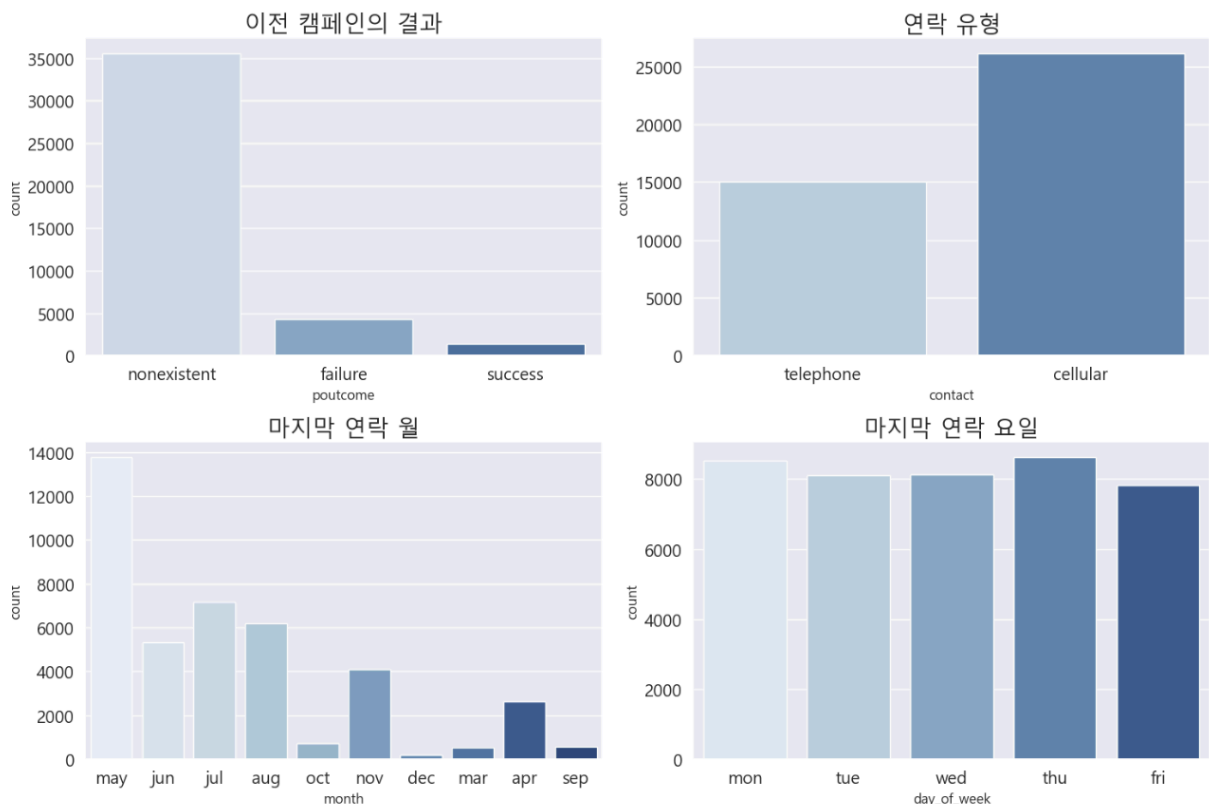
데이터에 존재하는 각 컬럼들이 어떤 타입인지, 범주형 데이터인 경우 어떤 값들이 존재하는지, 연속형 변수인 경우 변수의 분포가 어떻게 구성되었는지 확인한다.

### 4-1. 범주형 데이터 분포



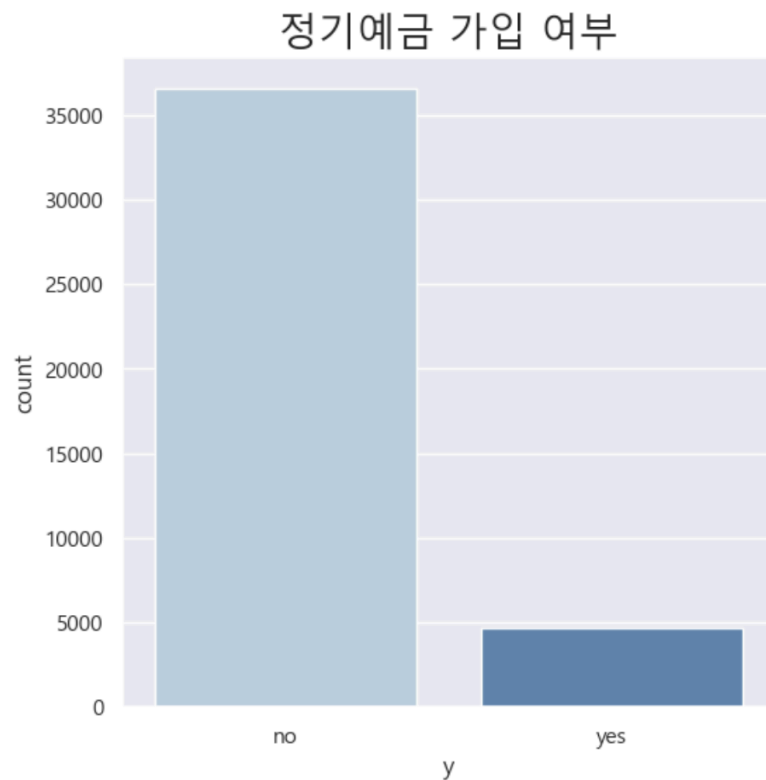
- 직업별 종사자 수
  - 주로 서비스직종이 많이 분포되었으며, 회사원의 비율은 중간정도 차지한다.
- 결혼 여부

- 결혼한 고객의 비율이 전체 데이터의 절반 이상을 차지하며, 아직 결혼하지 않은 고객의 비율이 결혼한 고객 다음으로 많다
- 최종 학력
  - 기초학력인 4학년이 최종학력인 고객의 수가 가장 많았으며, 고등학교까지 졸업한 고객이 그 다음으로 많았다.
- 주택 대출 여부
  - 주택 대출을 받지 않은 고객과, 주택을 대출받은 고객의 비율이 서로 비슷하다.



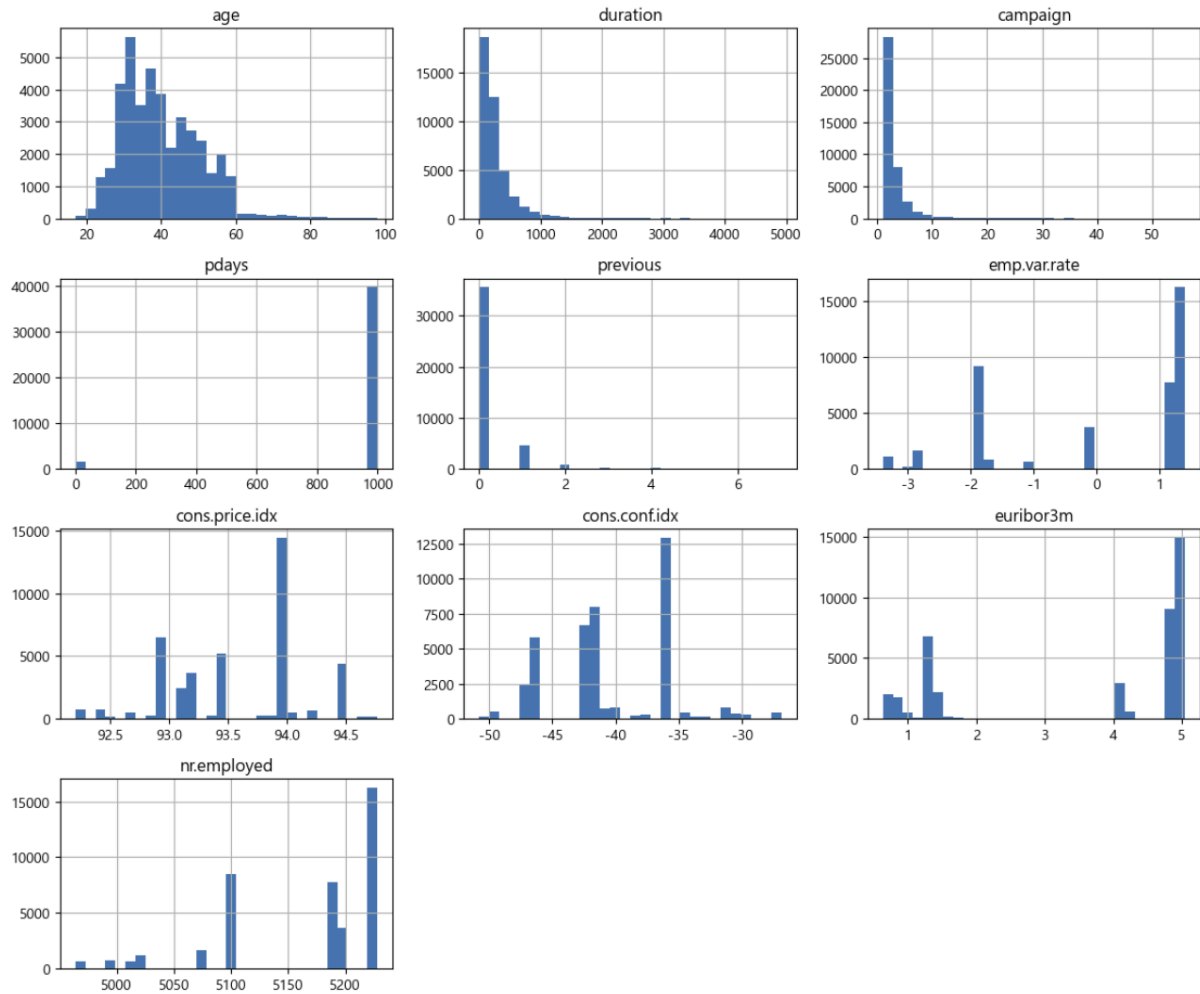
- 이전 캠페인의 결과
  - 직전 캠페인을 통해서 예금상품에 가입한 고객은 1,373명, 가입하지 않은 고객은 4,252명으로 약 5,500명 중에서 1/3이 직전 캠페인을 통해 예금상품에 가입하였다.
  - 이전 캠페인에 대상이 아닌 고객은 35,563명이다.
- 연락 유형
  - 유선 집 전화기(telephone) : 15,044명에게 유선 전화로 캠페인에 대해 설명하였다.
  - 휴대폰(cellular) : 26,144명에게 휴대폰으로 캠페인에 대해 설명하였다.
- 마지막 연락 월
  - 5월달에 마지막으로 은행과 연락한 고객이 가장 많았다.

- 마지막 연락 요일
  - 주말을 제외한 모든 요일에 고르게 분포되어 있다.



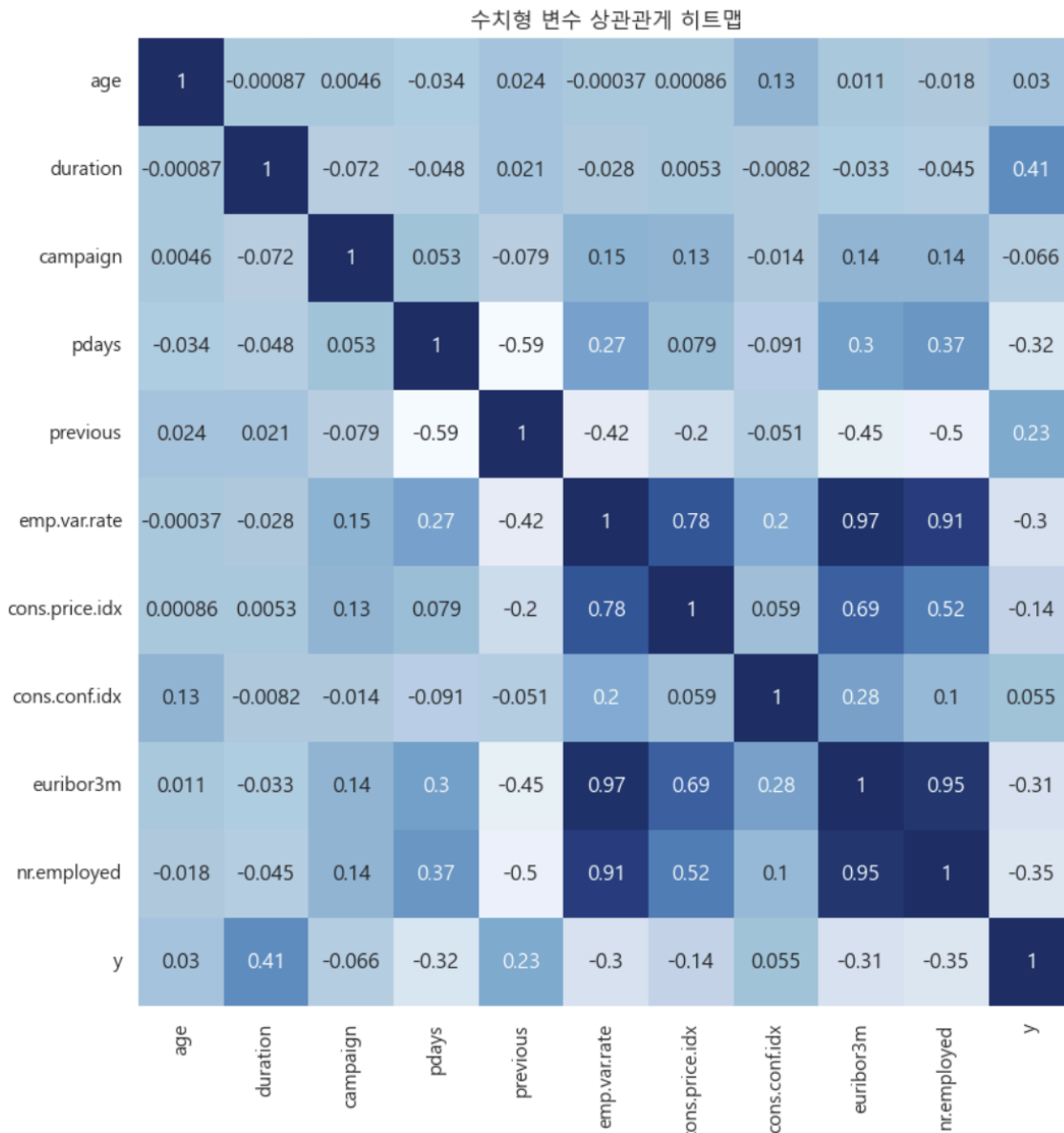
- 정기예금 가입 여부
  - 캠페인을 통해 정기예금에 가입한 고객은 약 4,640명이 가입하였으며, 36,548명은 가입하지 않았다.
  - 해당 컬럼을 타겟변수로 설정하여 모델을 생성한다.

## 4-2. 연속형 데이터 분포



- **age** : 20대 ~ 60대 사이에 주로 분포되었으며, 정규분포 형상을 확인할 수 있다.
- **duration** : 통화 시간을 나타내며, 통화시간이 짧은 경우가 많다.
- **campaign** : 캠페인 기간 동안 통화를 한 횟수를 나타내며, 대부분 0에 가깝다.
- **pdays** : 이전 캠페인 후 지난 일수를 나타내며, 대부분의 고객이 999일에 분포되어있다. 해당 데이터는 pdays가 27일이 경과되었을 경우 전부 999일로 변환되어있다.
- **previous** : 이전 캠페인 동안 연락 횟수를 나타내며, 대부분 0~1 사이에 분포되어 있다.
- **emp.var.rate** : 고용 변동률은 일정 주기마다 변하는 수치로 경제 지표와 비슷한 성향 가진다.
- **cons.price.idx** : 소비자 물가지수 또한 지속적으로 변경되는 수치로, emp.var.rate와 같이 경제 지표와 비슷한 성향을 가진다.
- **cons.conf.idx** : 소비자 신뢰지수 또한 지속적으로 변경되는 수치로, emp.var.rate와 같이 경제 지표와 비슷한 성향을 가진다.
- **euribor3m** : 3개월 유리보 금리는 유럽 국가 간의 금리를 뜻하며, 해당 변수 또한 변동성이 존재하는 변수에 속하며, emp.var.rate와 같이 경제 지표와 비슷한 성향을 가진다.
- **nr.employed** : 고용자 수 데이터 또한 지속적으로 변경되는 수치로, emp.var.rate와 같이 경제 지표와 비슷한 성향을 가진다.

### 4-3. 연속형 변수와 정기에금 가입여부 상관관계



정기에금 가입 여부와 상관성이 높은 연속형 변수 top5

1. duration : 0.41(양의 상관관계)
2. re.employed : -0.35(음의 상관관계)
3. pdays : -0.32(음의 상관관계)
4. euribor3m : -0.31(음의 상관관계)
5. emp.var.rate : -0.3(음의 상관관계)

### 4-4. 정기에금 가입 여부에 따른 수치형 데이터 평균값 비교

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
y										
0	39.911185	220.844807	2.633085	984.113878	0.132374	0.248875	93.603757	-40.593097	3.811491	5176.166600
1	40.913147	553.191164	2.051724	792.035560	0.492672	-1.233448	93.354386	-39.789784	2.123135	5095.115991

- age : 정기예금 가입 여부는 나이와 큰 차이가 없다.
- duration : 예금을 가입한 사람들은 평균 통화 지속시간이 길다.
- campaign : 캠페인을 진행하면서 연락한 횟수는 큰 차이가 없으나, 가입하지 않은 사람들의 평균 전화 횟수가 0.6회정도 더 많다.
- pdays : 예금을 가입한 고객들의 최근 연락날짜는 예금을 가입하지 않은 고객보다 기간이 짧다.
- previous : 예금을 가입한 고객들의 평균적으로 이전 캠페인동안 연락한 횟수가 높다
- emp.var.rate : 예금을 가입한 당시의 고객들의 평균 고용 변동률이 낮다.
- cons.price.idx : 소비자 물가지수는 큰 차이가 없다.
- cons.conf.idx : 소비자 신뢰지수는 큰 차이가 없다.
- euribor3m : 예금을 가입한 고객들의 당시 3개월 유리보 금리는 평균적으로 예금을 가입하지 않은 고객보다 낮다.
- nr.employed : 예금을 가입한 고객들의 당시 고용자 수는 평균적으로 예금을 가입하지 않은 고객보다 낮다.

## 5. 데이터 전처리

### 5-1. 범주형 데이터 원-핫 인코딩

```
# 범주형 컬럼 리스트
cat_col = ['job', 'marital', 'education', 'default', 'housing', 'loan',
           'contact', 'month', 'day_of_week', 'poutcome']

# 범주형 컬럼만 원-핫 인코딩 진행
df_ohe_encoded = pd.get_dummies(df, columns=cat_col, drop_first=True)

# 원-핫 인코딩 완료된 데이터 확인
df_ohe_encoded.sample(5)
```

day_of_week_mon	day_of_week_thu	day_of_week_tue	day_of_week_wed	poutcome_nonexistent	poutcome_success
0	1	0	0	0	0
0	0	1	0	1	0

분류 모델 훈련에 필요한 데이터 셋의 범주형 데이터를 모델이 학습할 수 있도록 수치형 변수로 변환

## 5-2. 수치형 데이터 표준화

```
from sklearn.preprocessing import StandardScaler

# 표준화 함수 정의
scaler = StandardScaler()

# 표준화 진행할 컬럼 리스트 num_col
df_ohe_encoded[num_col] = scaler.fit_transform(df_ohe_encoded[num_col])

# 표준화 처리된 데이터프레임 별도 저장
df_scaled = df_ohe_encoded.copy()

df_scaled.sample(5)
```

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
19699	0.477486	1.985200	0.878132	0.195414	-0.349494	0.839061	-0.227465	0.951267	0.776458	0.845170
40810	-0.961898	0.565864	-0.565922	0.195414	5.712397	-0.752343	1.076883	0.648770	-1.581670	-2.815697

분류 모델의 성능 향상을 위해 수치형 변수들의 범위를 비슷하게 맞춰주도록 한다.

## 5-3. 훈련, 테스트 데이터 분리

```
from sklearn.model_selection import train_test_split

# 데이터 X, y 분리
X = df_scaled.drop(columns='y')
y = df_scaled[['y']]

# 훈련, 테스트셋 비율 7:3 으로 분
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# 훈련 데이터 크기
print(f"훈련 데이터 크기 : {X_train.shape}")

# 테스트 데이터 크기
print(f"테스트 데이터 크기 : {X_test.shape}")
```

- 훈련 데이터 크기 : (28,831, 53)
- 테스트 데이터 크기 : (12,357, 53)

## 6. 분류 모델 생성



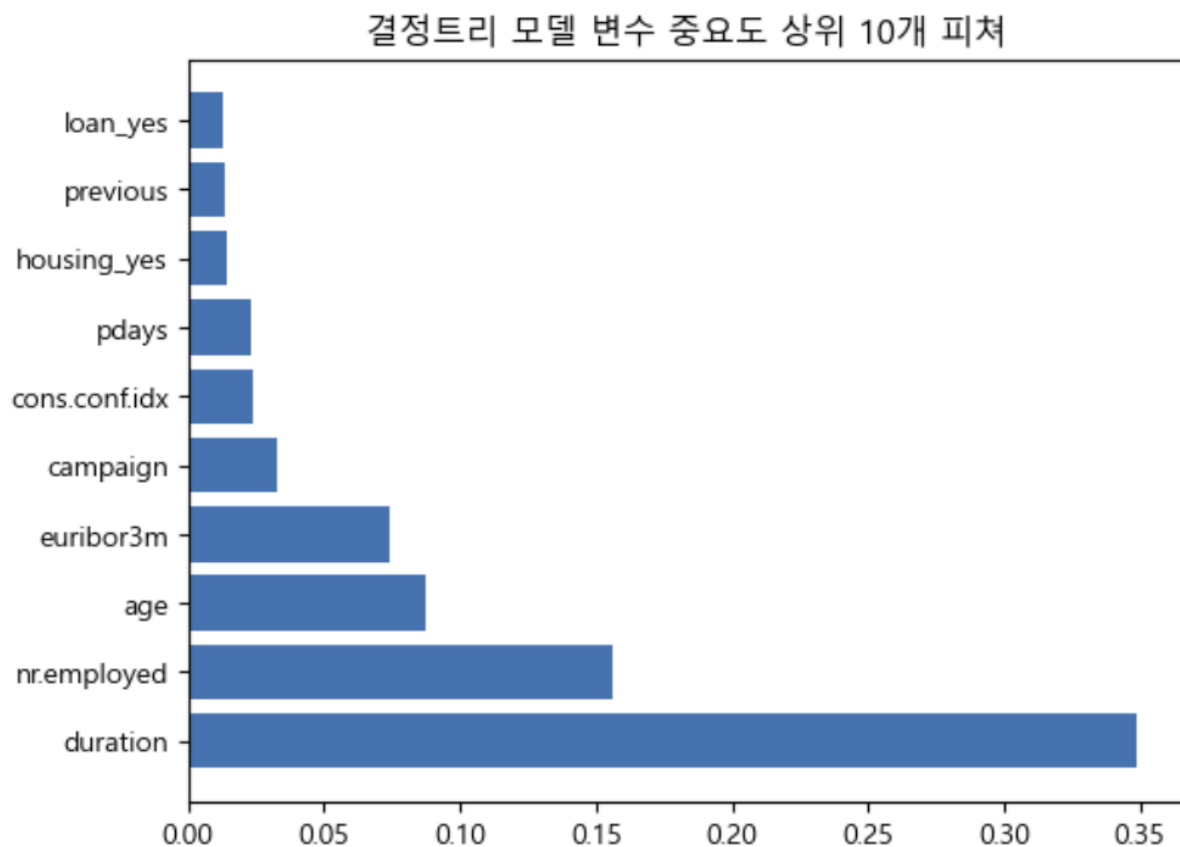
## 6-1. 모델 종류

- 로지스틱 회귀분석 : 명칭은 회귀분석이지만, 분류를 진행할 때 사용되는 모델로 사용된다. 선형데이터에 적용이 가능한 모델로, 해당 데이터셋에는 적용이 어려움
- 의사결정나무 : 트리 구조를 기반으로 의사결정을 수행하는 모델
  - Decision Tree
- 랜덤 포레스트 분류분석 : 여러 개의 의사결정나무를 결합하여 성능을 향상한 모델
  - RandomForest Classification
- 그래디언트 부스팅 : 여러 개의 약한 학습기를 결합하여 강한 분류기를 만드는 모델
  - XGBoost, LightGBM, CatBoost

## 6-2. 의사결정 나무

데이터 분류의 기본 모델로 활용되는 의사결정나무를 사용해서 데이터를 분류하고, 분류를 얼마나 정확하게 진행하였는지 점수를 확인

- 의사결정나무 분류 점수 : 0.8899점
- 의사결정나무 변수 중요도 상위 10개 변수

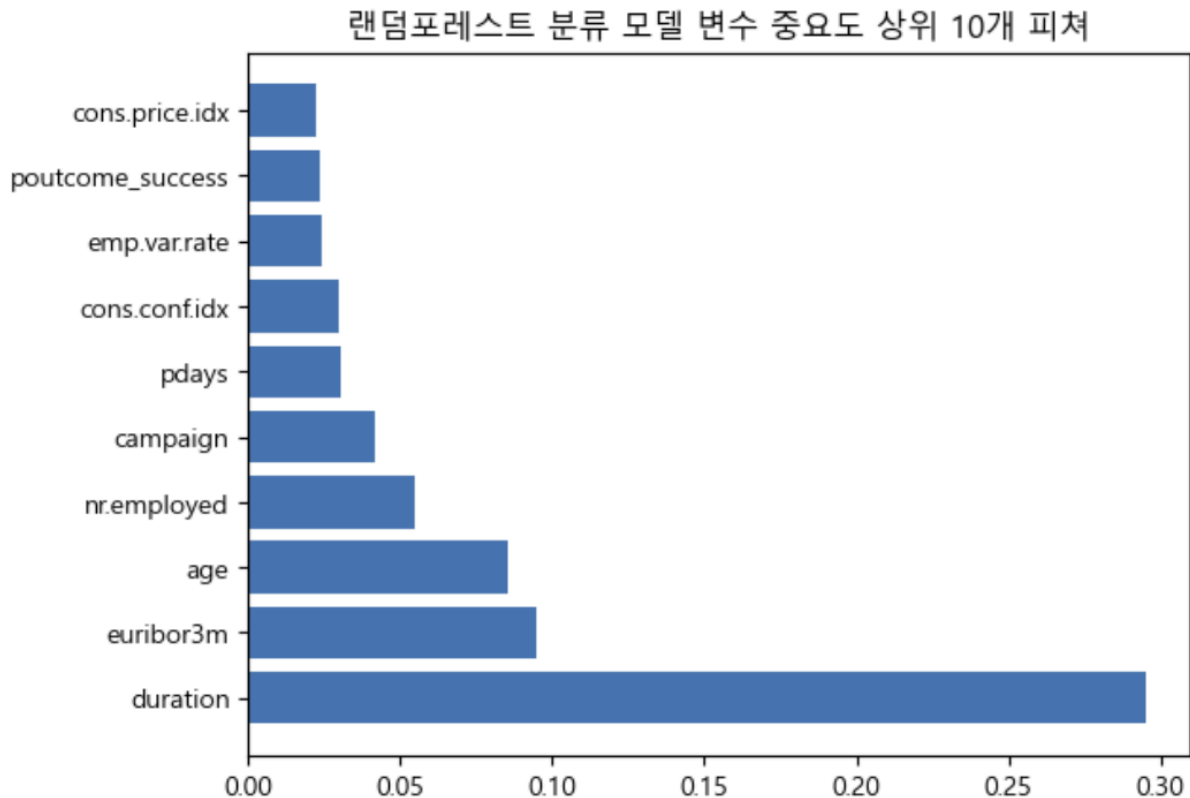


duration, nr.employed, age 변수들이 모델에서 높은 중요도를 차지하고 있다.

### 6-3. 랜덤포레스트 분류

여러 의사결정나무들이 앙상블된 랜덤포레스트 분류분석을 진행하고, 분류 정확도 점수를 확인하여 모델의 성능을 확인한다.

- 랜덤포레스트 분류 점수 : 0.9419점
- 랜덤포레스트 변수 중요도 상위 10개 변수



duration, euribor3m, age 변수들이 모델에서 높은 중요도를 차지하고 있다.

### 6-4. 분류모델 확인 결과

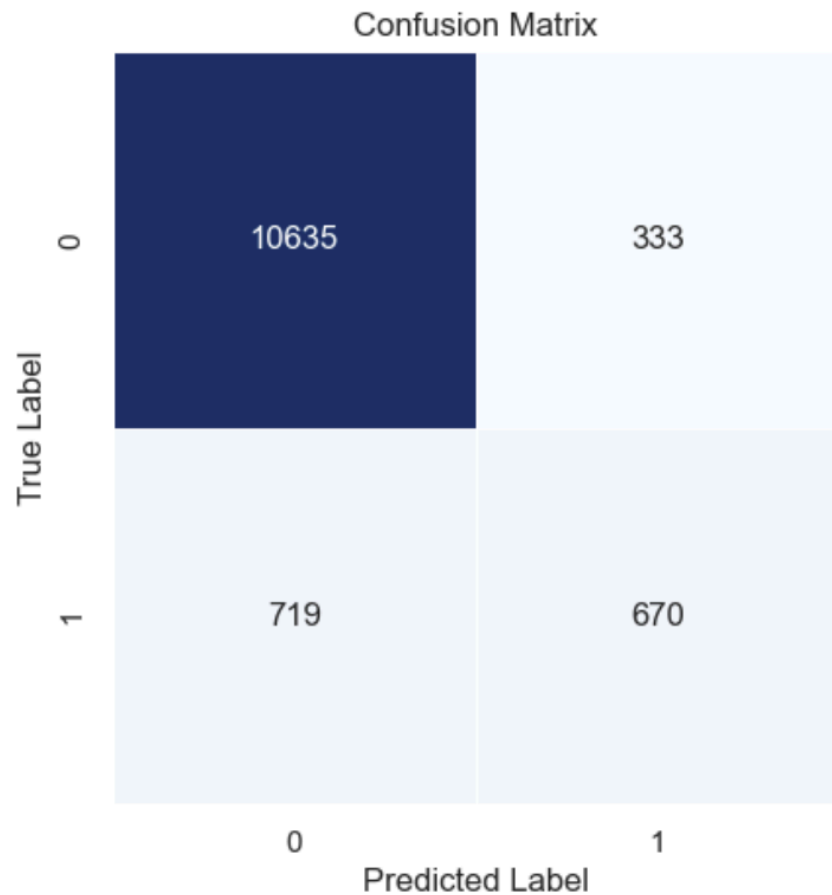
의사결정나무와 랜덤포레스트 분류 분석을 진행한 결과 두 모델 전부 정기에금에 가입하거나 가입하지 않을 고객을 잘 분류하는 것으로 확인된다.

두 모델 중에서 점수가 더 높은 랜덤포레스트 분류 모델에 대해 분류 정확도, 정밀도 등을 추가로 계산하고, 변수들을 통해 마케팅 전략을 도출한다.

## 7. 분류 정확도, 정밀도, 오분류율 산출

### 7-1. 혼동행렬 산출

혼동행렬은 생성된 분류 예측 모델이 옳게 예측했거나, 틀리게 예측했는지를 행렬로 표현하여, 모델의 정확도, 정밀도, 재현율, 오분류율 등을 확인할 수 있다.



랜덤포레스트 분류 분석을 통한 혼동행렬

- True Positive(TP) : 실제 참인 데이터 중에서 참이라고 맞게 예측한 데이터(10,635개)
- True Negative(TN) : 실제 참이 아닌 데이터 중에서 참이 아니라고 맞게 예측한 데이터(670개)
- False Positive(FP) : 실제로는 참이 아닌데, 참이라고 틀리게 예측한 데이터(333개)
- False Negative(FN) 실제로는 참인데, 참이 아니라고 틀리게 예측한 데이터(719개)

## 7-2. 성능 지표 산출

	Precision	Recall	F1-Score	support
0(예금 가입X)	0.94	0.97	0.95	10968
1(예금 가입O)	0.67	0.48	0.56	1389
accuracy			0.91	12357
macro avg	0.80	0.73	0.76	12357
weighted avg	0.91	0.91	0.91	12357

- 정확도(accuracy) : 91%의 정확도로 단순 수치만 비교할 경우 예측을 잘 하고 있다고 가정
- 정밀도(precision) :
  - 0 (예금에 가입하지 않은 고객) : 94%의 정밀도로 가입하지 않는다고 예측했을 때, 대부분 가입하지 않았다.

- 1 (예금에 가입한 고객) : 가입한다고 예측한 경우 중에서 67%가 실제로 가입하였다.
- 재현율(recall) :
  - 0 : 97% → 실제 가입하지 않는 고객을 잘 찾아냄.
  - 1 : 48% → 실제 가입할 고객 중 절반 이상을 놓치고 있음(재현율이 낮음).
- F1-score:
  - 0 : 0.95로 매우 우수.
  - 1 : 0.56로 낮음 → 가입할 고객을 잘 예측하지 못함.
- 평균 지표:
  - **macro avg** : 단순 평균 → 정밀도(0.80), 재현율(0.73), F1-score(0.76)
  - **weighted avg** : 데이터 비율을 반영한 평균 → 전체적으로 높은 이유는 0이 압도적으로 많기 때문.
- 오분류율 :
  - **error\_rate** : 8% 비율로 전체 데이터에서 잘못 예측하였다.

## 8. 마케팅 제안

### 8-1. 고객 상담 전략 강화 (duration)

마지막 통화 지속 시간이 모델에서 가장 중요한 피쳐로 나타남 → 고객과의 상담 시간이 증가할수록 고객이 정기 예금에 가입할 확률이 높음

 마케팅 전략 :

- 영업 상담원이 고객과 충분한 시간을 갖고 설명할 수 있도록 **대화 스크립트 및 상담 기술 강화**
- 기존 가입 고객의 주요 관심사를 분석하여 **맞춤형 상담 내용 제공**
- 상담 중 특정 반응(예: 관심 표시, 질문 빈도)이 있는 고객을 타겟으로 추가 컨택을 진행

### 8-2. 금리 혜택 홍보 (euribor3m)

3개월 유리보 금리가 낮을 때 정기 예금에 가입할 확률이 높음

 마케팅 전략 :

- 금리 변동 실시간 모니터링을 통해 금리가 낮은 시점에 예금 가입 홍보 캠페인을 적극적으로 진행
- 낮은 금리와 관련된 홍보문구를 활용하여 고객의 관심도를 가져옴

### 8-3. 연령별 맞춤 마케팅 (age)

특정 연령대별로 맞춤 혜택을 제공

 마케팅 전략 :

- 청년층(20~30대) : 단기 고금리 상품, 월급 계좌를 자사 은행으로 옮길 경우 예금 이자 혜택 제공
- 중장년층(40대~60대) : 안정적인 장기 예금 상품 마련, 노후 대비 은퇴자금을 마련할 수 있도록 금융 컨설팅 제공 및 예금 상품 추천

#### 8-4. 경제 지표 (nr.employed, emp.var.rate, cons.price.idx, cons.conf.idx)


사회 경제 지표와 관련된 변수들은 직접 통제할 수 없는 변수들로 시기에 따라서 유동적인 마케팅 방법을 적용하여 고객의 예금 가입 전환을 위한 전략을 수립

 마케팅 전략 :

- 경제 불안정 시기 : 고객들에게 불안한 경제 상황을 설명함과 동시에 안전한 투자 상품인 저금리 예금 상품을 추천
- 경제 성장기 : 손실 위험이 크지만 이자가 높은(하이리스크 하이리턴) 상품을 적극 추천하여 경제 성장기에 맞춰 높은 이자를 받아갈 수 있도록 홍보함

#### 8-5. 캠페인 전략 (campaign, pdays, poutcome\_success)

이전 캠페인 결과를 바탕으로 예금을 가입한 고객들과 가입하지 않은 고객을 비교

 마케팅 전략 :

- 너무 잦은 연락을 지양 : 예금 가입 권유 전화를 자주 할 수록 고객의 관심도는 낮아질 수 있음
- pdays가 999일로 분류된 고객들에게 주로 연락을 진행