산탄데르 고객 만족 예측 - 분류

학습 내용

- LightGBM 모델을 활용한 예측

데이터 설명

- 데이터 다운로드 : https://www.kaggle.com/c/santander-customer-satisfaction/data (https://www.kaggle.com/c/santander-customer-satisfaction/data)
- 370개의 피처로 이루어진 데이터
- 피처 이름은 전부 익명처리되어 있음.
- 클래스 레이블 명은 TARGET
  - 값이 1이면 불만을 가지고 있음.
  - 값이 0이면 만족한 고객

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib
```

In [2]:

```
train = pd.read_csv("../../data/santander_customer/train.csv", encoding='latin-1')
test = pd.read_csv("../../data/santander_customer/test.csv", encoding='latin-1')
sub = pd.read_csv("../../data/santander_customer/sample_submission.csv")

train.shape, test.shape, sub.shape
```

Out[2]:

```
((76020, 371), (75818, 370), (75818, 2))
```

In [3]:

```
## ID 열을 삭제
# train.drop('ID', axis=1, inplace=True)
train = train.loc[  :, "var3":  ]
train.head()
```

Out[3]:

	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_comer_ult3	imp_o
0	2	23	0.0	0.0	0.0	
1	2	34	0.0	0.0	0.0	
2	2	23	0.0	0.0	0.0	
3	2	37	0.0	195.0	195.0	
4	2	39	0.0	0.0	0.0	

5 rows × 370 columns

In [4]:

```
# 피처와 레이블를 지정.
X = train.iloc[:, :-1]
y = train['TARGET']

from sklearn.model_selection import train_test_split

X_train , X_test, y_train, y_test = train_test_split(X, y,
                                                     test_size=0.2, random_state=0)

X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

Out[4]:

```
((60816, 369), (15204, 369), (60816,), (15204,))
```

In [9]:

```
from lightgbm import LGBMClassifier
from sklearn.metrics import roc_auc_score
```

In [10]:

```
%%time

lgbm_model = LGBMClassifier(n_estimators = 500)
evals = [(X_test, y_test)]

lgbm_model.fit(X_train, y_train, early_stopping_rounds=100,
               eval_metric='auc', eval_set=evals,
               verbose=True)
```

```
[87]    valid_0's auc: 0.836538 valid_0's binary_logloss: 0.14005
[88]    valid_0's auc: 0.836583 valid_0's binary_logloss: 0.140073
[89]    valid_0's auc: 0.836427 valid_0's binary_logloss: 0.140128
[90]    valid_0's auc: 0.836458 valid_0's binary_logloss: 0.140113
[91]    valid_0's auc: 0.836471 valid_0's binary_logloss: 0.140151
[92]    valid_0's auc: 0.836582 valid_0's binary_logloss: 0.140107
[93]    valid_0's auc: 0.836317 valid_0's binary_logloss: 0.140177
[94]    valid_0's auc: 0.836218 valid_0's binary_logloss: 0.140221
[95]    valid_0's auc: 0.836338 valid_0's binary_logloss: 0.140177
[96]    valid_0's auc: 0.836151 valid_0's binary_logloss: 0.140256
[97]    valid_0's auc: 0.836344 valid_0's binary_logloss: 0.140245
[98]    valid_0's auc: 0.836296 valid_0's binary_logloss: 0.14029
[99]    valid_0's auc: 0.836433 valid_0's binary_logloss: 0.140272
[100]   valid_0's auc: 0.836407 valid_0's binary_logloss: 0.140291
[101]   valid_0's auc: 0.836355 valid_0's binary_logloss: 0.140319
[102]   valid_0's auc: 0.836324 valid_0's binary_logloss: 0.140326
[103]   valid_0's auc: 0.836208 valid_0's binary_logloss: 0.140332
[104]   valid_0's auc: 0.836412 valid_0's binary_logloss: 0.140295
[105]   valid_0's auc: 0.836649 valid_0's binary_logloss: 0.140254
[106]   valid_0's auc: 0.836643 valid_0's binary_logloss: 0.140268
```

- 수행 시간이 상당히 줄어들었음.

In [11]:

```
pred_prob = lgbm_model.predict_proba(X_test)[:, 1]
lgbm_roc_score = roc_auc_score(y_test, pred_prob, average='macro')
print("ROC AUC : {0:.4f}".format(lgbm_roc_score))
```

```
ROC AUC : 0.8408
```

In [12]:

```
%%time

from sklearn.model_selection import GridSearchCV

lgbm_model01 = LGBMClassifier(n_estimators = 500)

params = {"max_depth":[128,160],
          "min_child_samples":[60, 100],
          "num_leaves":[32, 64],
          "subsamle":[0.8, 1]}

gridcv = GridSearchCV(lgbm_model01, param_grid=params, cv=3)
gridcv.fit(X_train, y_train, early_stopping_rounds=30,
           eval_metric='auc',
           eval_set = [(X_train, y_train), (X_test, y_test)])
```

```
[63]    training's auc: 0.933838        training's binary_logloss: 0.102386     vali
d_1's auc: 0.839297     valid_1's binary_logloss: 0.139429
[64]    training's auc: 0.934193        training's binary_logloss: 0.102083     vali
d_1's auc: 0.838952     valid_1's binary_logloss: 0.139477
[65]    training's auc: 0.934873        training's binary_logloss: 0.101827     vali
d_1's auc: 0.83894      valid_1's binary_logloss: 0.139477
[66]    training's auc: 0.935675        training's binary_logloss: 0.101485     vali
d_1's auc: 0.838836     valid_1's binary_logloss: 0.139533


[67]    training's auc: 0.936051        training's binary_logloss: 0.101253     vali
d_1's auc: 0.83869      valid_1's binary_logloss: 0.139574
Wall time: 1min 51s
```

Out[12]:

```
GridSearchCV(cv=3, estimator=LGBMClassifier(n_estimators=500),
             param_grid={'max_depth': [128, 160],
                         'min_child_samples': [60, 100], 'num_leaves': [32, 64],
                         'subsamle': [0.8, 1]})
```

In [14]:

```
print("GridSearchCV  최적 파라미터 : ", gridcv.best_params_ )
```

```
GridSearchCV  최적 파라미터 :  {'max_depth': 128, 'min_child_samples': 60, 'num_leav
es': 64, 'subsamle': 0.8}
```

In [15]:

```
pred_prob = gridcv.predict_proba(X_test)[:, 1]
lgbm_roc_score = roc_auc_score(y_test, pred_prob, average='macro')
print("ROC AUC : {0:4f}".format(lgbm_roc_score))
```

```
ROC AUC : 0.841443
```

최종 모델

In [18]:

```
%%time

lgbm_model_l = LGBMClassifier(n_estimators=1000,
                              max_depth=128,
                              min_child_samples=60,
                              num_leaves=64,
                              subsamle=0.8)

evals = [(X_test, y_test)]
lgbm_model_l.fit(X_train, y_train, early_stopping_rounds=100,
                 eval_metric='auc', eval_set=evals,
                 verbose=True)
```

```
C:₩Users₩toto₩anaconda3₩lib₩site-packages₩lightgbm₩sklearn.py:726: UserWarning: 'ear
ly_stopping_rounds' argument is deprecated and will be removed in a future release o
f LightGBM. Pass 'early_stopping()' callback via 'callbacks' argument instead.
  _log_warning("'early_stopping_rounds' argument is deprecated and will be removed i
n a future release of LightGBM. "
C:₩Users₩toto₩anaconda3₩lib₩site-packages₩lightgbm₩sklearn.py:736: UserWarning: 'ver
bose' argument is deprecated and will be removed in a future release of LightGBM. Pa
ss 'log_evaluation()' callback via 'callbacks' argument instead.
  _log_warning("'verbose' argument is deprecated and will be removed in a future rel
ease of LightGBM. "
```

In [19]:

```
pred_prob = lgbm_model_l.predict_proba(X_test)[:, 1]
lgbm_roc_score = roc_auc_score(y_test, pred_prob, average='macro')
print("ROC AUC : {0:4f}".format(lgbm_roc_score))
```

```
ROC AUC : 0.841443
```

In [ ]: