In [1]:
```
import pandas as pd
import numpy as np
from scipy import stats
import math

# matplotlit and seaborn for visualizations
import seaborn as sns
import matplotlib.pyplot as plt
plt.rcParams['font.size'] = 12

# Suppress warnings from pandas
import warnings
warnings.filterwarnings('ignore')

# modeling
import lightgbm as lgb

# utilities
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_auc_score

sns.set_style("darkgrid")
# memory management
import gc
import os
# 운영체제별 한글 폰트 설정
if os.name == 'posix': # Mac 환경 폰트 설정
    plt.rc('font', family='AppleGothic')
elif os.name == 'nt': # Windows 환경 폰트 설정
    plt.rc('font', family='Malgun Gothic')

plt.rc('axes', unicode_minus=False) # 마이너스 폰트 설정

# 글씨 선명하게 출력하는 설정
%config InlineBackend.figure_format = 'retina'

df = pd.read_csv("dataset2.csv")
# import mglearn
# from sklearn.model_selection import KFold
# from sklearn.preprocessing import LabelEncoder
```

In [2]:
```
len(df[df['당뇨여부']==1])
```

Out[2]:
```
46579
```

In [3]:
```
nodang = df[df['당뇨여부']==0].copy()
dang = df[df['당뇨여부']==1].copy()
print(len(nodang))
print(len(dang))
```

```
1133541
46579
```

In [4]:
```
X = df[['허리둘레','연령대코드(5세단위)', '감마지티피']].copy()
y = df[['당뇨여부']].copy()
```

비당뇨 : 당뇨 = 10 : 1

In [5]:
```
from imblearn.under_sampling import RandomUnderSampler

rus = RandomUnderSampler(random_state=42, sampling_strategy=0.1)
X_rus, y_rus = rus.fit_resample(X, y)
```

In [6]:
```
print(len(X_rus))
print(len(y_rus))
print(np.sum(y_rus))
```

```
512369
512369
당뇨여부    46579.0
dtype: float64
```

In [7]:
```
y_rus
```

Out[7]:

	당뇨여부
0	0.0
1	0.0
2	0.0
3	0.0
4	0.0
...	...
512364	1.0
512365	1.0
512366	1.0
512367	1.0
512368	1.0

512369 rows × 1 columns

In [8]:
```
X_train, X_test, y_train, y_test = train_test_split(X_rus,y_rus.values,strati
from lightgbm import LGBMClassifier, plot_importance
from sklearn.preprocessing import StandardScaler, RobustScaler
sc = RobustScaler()
x_train = sc.fit_transform(X_train)
x_test = sc.transform(X_test)
lgb = LGBMClassifier(n_estimators=400)
evals = [(x_test, y_test)]
```

In [9]:
```
# lgb.fit(x_train, y_train, early_stopping_rounds=100, eval_metric="logloss",
lgb.fit(x_train, y_train, eval_metric="logloss",eval_set=evals, verbose=False
```

Out[9]:
```
LGBMClassifier(n_estimators=400)
```

In [10]:
```
y_pred = lgb.predict(x_test)
```

정확도

In [11]:
```
lgb.score(x_test, y_test)
```

Out[11]:
```
0.9090793762320198
```

예측 결과

In [12]:
```
print("당뇨로 예측한 데이터 수 :", np.sum(y_pred))
print("실제 당뇨 데이터 수 :", np.sum(y_test))
print("전체 데이터 수 :", len(y_test))
```

```
당뇨로 예측한 데이터 수 : 33.0
실제 당뇨 데이터 수 : 9316.0
전체 데이터 수 : 102474
```

TN, FP, FN, TP

In [13]:
```
from sklearn import metrics
metrics.confusion_matrix(y_test, y_pred)

# [[TN,FP],
#  [FN,TP]]
```

Out[13]:
```
array([[93141,    17],
       [ 9300,    16]])
```

TN, FP, FN, TP 검토

In [14]:
```
y_test = y_test.reshape(-1)
```

In [15]:
```
P = sum(y_test)
TP = sum((y_test==1) & (y_pred==1))
TPR = TP/P
FN = sum((y_test==1) & (y_pred==0))
FNR = FN/P
N = sum(y_test==0)
TN = sum((y_test==0) & (y_pred==0))
TNR = TN/N
FP = sum((y_test==0) & (y_pred==1))
FPR = FP/N
print(TN, FP, FN, TP)
print(TPR, FPR)
```

```
93141 17 9300 16
0.0017174753112924003 0.00018248566950771806
```

데이터 별 당뇨 분류 확률

In [16]:
```
y_pred_proba = lgb.predict_proba(x_test)
y_pred_proba
```

Out[16]:
```
array([[0.99415624, 0.00584376],
       [0.9751098 , 0.0248902 ],
       [0.91464609, 0.08535391],
       ...,
       [0.93888085, 0.06111915],
```

```
       [0.96920184, 0.03079816],
       [0.89094515, 0.10905485]])
```

[당뇨X 확률, 당뇨 확률]

In [17]:
```
y_pred_proba.shape
```

Out[17]:
```
(102474, 2)
```

데이터 별 당뇨로 예측할 확률

In [18]:
```
pos_proba = y_pred_proba[:,1]
# np.unique(pos_proba)
print(len(pos_proba[pos_proba > 0.8]))
```

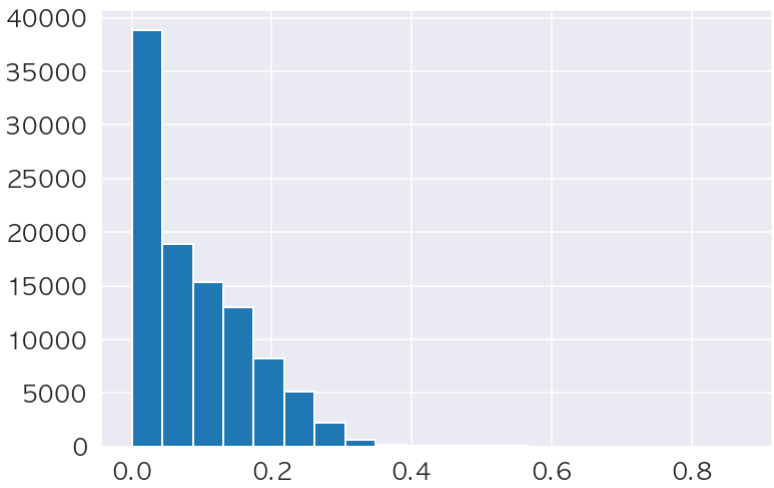
```
3
```

In [19]:
```
print("pos_proba 최솟값 :", min(pos_proba))
print("pos_proba 최댓값 :", max(pos_proba))
```

```
pos_proba 최솟값 : 1.3247325044621881e-06
pos_proba 최댓값 : 0.8713003864097405
```

In [20]:
```
plt.hist(pos_proba, range=(min(pos_proba),max(pos_proba)), bins=20)
```

Out[20]:
```
(array([3.8812e+04, 1.8936e+04, 1.5337e+04, 1.2985e+04, 8.2040e+03,
        5.0760e+03, 2.2110e+03, 6.1600e+02, 1.6500e+02, 6.2000e+01,
        2.8000e+01, 1.6000e+01, 1.3000e+01, 3.0000e+00, 3.0000e+00,
        1.0000e+00, 2.0000e+00, 1.0000e+00, 1.0000e+00, 2.0000e+00]),
 array([1.32473250e-06, 4.35662778e-02, 8.71312309e-02, 1.30696184e-01,
        1.74261137e-01, 2.17826090e-01, 2.61391043e-01, 3.04955996e-01,
        3.48520949e-01, 3.92085902e-01, 4.35650856e-01, 4.79215809e-01,
        5.22780762e-01, 5.66345715e-01, 6.09910668e-01, 6.53475621e-01,
        6.97040574e-01, 7.40605527e-01, 7.84170480e-01, 8.27735433e-01,
        8.71300386e-01]),
 <BarContainer object of 20 artists>)
```



당뇨 예측 확률과 실제 데이터

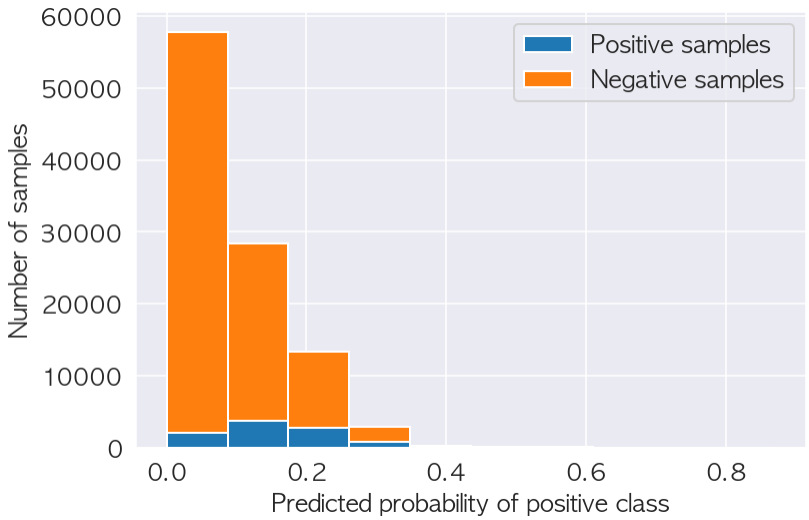
In [21]:
```
pos_sample_pos_proba = pos_proba[y_test==1]
neg_sample_pos_proba = pos_proba[y_test==0]
```

In [22]:
```

```

```
plt.hist([pos_sample_pos_proba, neg_sample_pos_proba], histtype="barstacked")
plt.legend(["Positive samples", "Negative samples"])
plt.xlabel("Predicted probability of positive class")
plt.ylabel("Number of samples")
```

Out[22]: Text(0, 0.5, 'Number of samples')

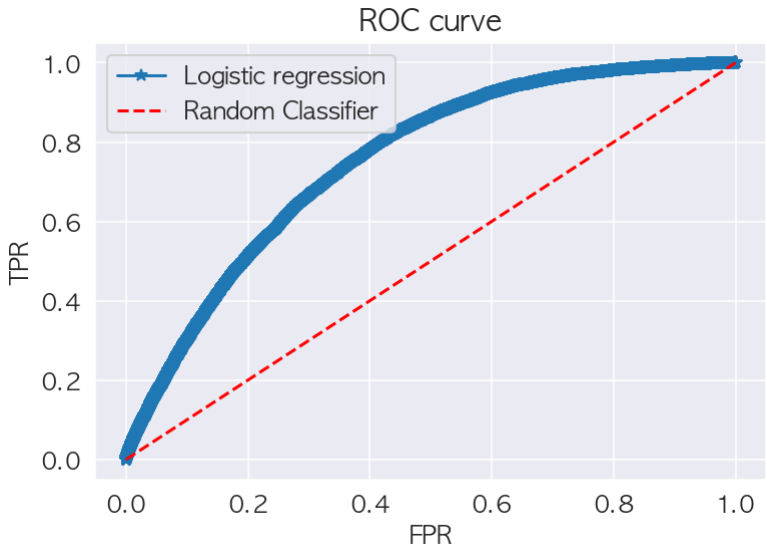


ROC Curve

In [23]:
```
fpr, tpr, thresholds = metrics.roc_curve(y_test, pos_proba)
```

In [24]:
```
plt.plot(fpr, tpr, '*-')
plt.plot([0,1], [0,1], 'r--')
plt.legend(['Logistic regression', 'Random Classifier'])
plt.xlabel('FPR')
plt.ylabel('TPR')
plt.title('ROC curve')
```

Out[24]: Text(0.5, 1.0, 'ROC curve')



AUC Score

In [25]:
```
metrics.roc_auc_score(y_test, pos_proba)
```

Out[25]: 0.7527034466326921

Precision, Recall, Accuracy

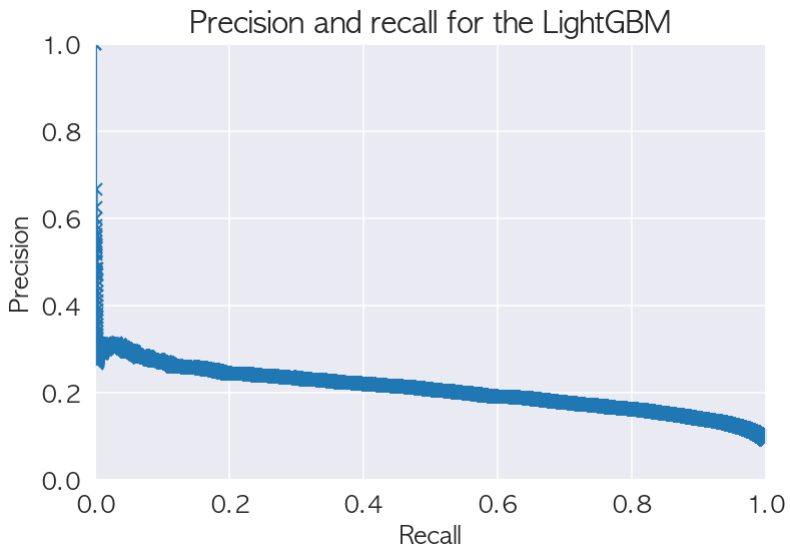
In [26]:
```
precision = TP/(TP+FP)
recall = TP/(TP+FN)
accuracy = (TP+TN)/(TP+FP+TN+FN)
print("precision :",precision)
print("recall :", recall)
print("accuracy :", accuracy)
```

```
precision : 0.48484848484848486
recall : 0.0017174753112924003
accuracy : 0.9090793762320198
```

In [27]:
```
precision2, recall2, thresholds2 = metrics.precision_recall_curve(y_test, pos
```

In [28]:
```
plt.plot(recall2, precision2, '-x')
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.title('Precision and recall for the LightGBM')
plt.xlim([0,1])
plt.ylim([0,1])
```

Out[28]: (0.0, 1.0)



In [29]:
```
metrics.auc(recall2, precision2)
```

Out[29]: 0.20555646303340713

In [ ]: