Performance Baselines

In [1]:
```
%reload_ext watermark
```

In [2]:
```
%load_ext watermark
%watermark -p scikit-learn,mlxtend,xgboost
```

```
The watermark extension is already loaded. To reload it, use:
  %reload_ext watermark
scikit-learn: 1.0.1
mlxtend     : 0.19.0
xgboost     : 1.5.0
```

Dataset

Source: https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset (https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset)

In [3]:
```
import pandas as pd


X_train = pd.read_csv('https://raw.githubusercontent.com/rasbt/stat451-machine-learning-fs21/main/hw02-starter/dataset/X_train.csv', header=None).values
y_train = pd.read_csv('https://raw.githubusercontent.com/rasbt/stat451-machine-learning-fs21/main/hw02-starter/dataset/y_train.csv', header=None).values.ravel().astype(int)

X_test = pd.read_csv('https://raw.githubusercontent.com/rasbt/stat451-machine-learning-fs21/main/hw02-starter/dataset/X_test.csv', header=None).values
y_test = pd.read_csv('https://raw.githubusercontent.com/rasbt/stat451-machine-learning-fs21/main/hw02-starter/dataset/y_test.csv', header=None).values.ravel().astype(int)

print('X_train.shape:', X_train.shape)
print('y_train.shape:', y_train.shape)
print('X_test.shape:', X_test.shape)
print('y_test.shape:', y_test.shape)
```

```
X_train.shape: (9119, 16)
y_train.shape: (9119,)
X_test.shape: (4492, 16)
y_test.shape: (4492,)
```

In [4]:
```
from sklearn import model_selection
from sklearn.model_selection import train_test_split


X_train_sub, X_valid, y_train_sub, y_valid = \
    train_test_split(X_train, y_train, test_size=0.2, random_state=1, stratify=y_train)

print('Train/Valid/Test sizes:', y_train.shape[0], y_valid.shape[0], y_test.shape[0])
```

```
Train/Valid/Test sizes: 9119 1824 4492
```

Baselines

Compare hyperparameter settings on validation set:

In [5]:
```
from sklearn.neighbors import KNeighborsClassifier


knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train_sub, y_train_sub)
print(f"Train Accuracy: {knn.score(X_train_sub, y_train_sub)*100:0.3f}%")
print(f"Valid Accuracy: {knn.score(X_valid, y_valid)*100:0.3f}%")
```

```
Train Accuracy: 79.657%
Valid Accuracy: 71.162%
```

In [6]:
```
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train_sub, y_train_sub)
print(f"Train Accuracy: {knn.score(X_train_sub, y_train_sub)*100:0.3f}%")
print(f"Valid Accuracy: {knn.score(X_valid, y_valid)*100:0.3f}%")
```

```
Train Accuracy: 84.003%
Valid Accuracy: 71.930%
```

In [7]:
```
knn = KNeighborsClassifier(n_neighbors=7)
knn.fit(X_train_sub, y_train_sub)
print(f"Train Accuracy: {knn.score(X_train_sub, y_train_sub)*100:0.3f}%")
print(f"Valid Accuracy: {knn.score(X_valid, y_valid)*100:0.3f}%")
```

```
Train Accuracy: 77.478%
Valid Accuracy: 69.518%
```

Choose best model and train on whole training set:

In [8]:
```
model = KNeighborsClassifier(n_neighbors=3)
model.fit(X_train, y_train)
print(f"Train Accuracy: {model.score(X_train, y_train)*100:0.3f}%")
print(f"Test Accuracy: {model.score(X_test, y_test)*100:0.3f}%")
```

```
Train Accuracy: 84.965%
Test Accuracy: 71.305%
```

In [11]:
```
# 4.3

from optuna.integration import LightGBMPruningCallback
import numpy as np
import lightgbm
import optuna

from sklearn.metrics import log_loss
from sklearn.model_selection import StratifiedKFold

import warnings

warnings.filterwarnings("ignore", category=UserWarning)
#optuna.logging.set_verbosity(optuna.logging.WARNING)


def objective(trial, X_train, y_train, cv=5):

    param_grid = {
        "n_estimators": trial.suggest_categorical("n_estimators", [10, 100]),
        "learning_rate": trial.suggest_categorical("learning_rate", [0.01]),
    }

    cv_iterator = StratifiedKFold(n_splits=cv, shuffle=True, random_state=123)

    cv_scores = np.zeros(cv)
    for idx, (train_sub_idx, valid_idx) in enumerate(cv_iterator.split(X_train, y_train)):

        X_train_sub, X_valid = X_train[train_sub_idx], X_train[valid_idx]
        y_train_sub, y_valid = y_train[train_sub_idx], y_train[valid_idx]

        model = lightgbm.LGBMClassifier(objective="multi_logloss", **param_grid)
        model.fit(
            X_train_sub,
            y_train_sub,
            eval_set=[(X_valid, y_valid)],
            eval_metric="multi_logloss",
            verbose=-1,
            early_stopping_rounds=50,
            callbacks=[
                LightGBMPruningCallback(trial=trial, metric="multi_logloss")
            ],  # Add a pruning callback to eliminate unpromising candidates
```

```
        )
        preds = model.score(X_valid, y_valid)

        cv_scores[idx] = preds

    return 1-np.mean(cv_scores)
```

In [12]:
```
study = optuna.create_study(direction="minimize", study_name="LGBM Classifier")

def func(trial):
    return objective(trial, X_train, y_train)

study.optimize(func, n_trials=50);
```

```
[I 2021-11-10 22:17:58,256] A new study created in memory with name: LGBM Classifier
[I 2021-11-10 22:18:00,553] Trial 0 finished with value: 0.07511800964286763 and parameters: {'n_est
imators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:00,832] Trial 1 finished with value: 0.48239960158212314 and parameters: {'n_est
imators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:01,109] Trial 2 finished with value: 0.48239960158212314 and parameters: {'n_est
imators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:01,384] Trial 3 finished with value: 0.48239960158212314 and parameters: {'n_est
imators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:01,682] Trial 4 finished with value: 0.48239960158212314 and parameters: {'n_est
imators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:01,970] Trial 5 finished with value: 0.48239960158212314 and parameters: {'n_est
imators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:02,257] Trial 6 finished with value: 0.48239960158212314 and parameters: {'n_est
imators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:02,557] Trial 7 finished with value: 0.48239960158212314 and parameters: {'n_est
imators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:05,341] Trial 8 finished with value: 0.07511800964286763 and parameters: {'n_est
imators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:05,640] Trial 9 finished with value: 0.48239960158212314 and parameters: {'n_est
imators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:08,097] Trial 10 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:10,777] Trial 11 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:13,354] Trial 12 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:15,883] Trial 13 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:18,448] Trial 14 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:21,054] Trial 15 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:23,817] Trial 16 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:26,408] Trial 17 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:28,988] Trial 18 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:31,599] Trial 19 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:34,782] Trial 20 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
```

```
[I 2021-11-10 22:18:37,797] Trial 21 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:40,443] Trial 22 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:43,170] Trial 23 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:45,837] Trial 24 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:48,489] Trial 25 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:51,131] Trial 26 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:53,963] Trial 27 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:56,682] Trial 28 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:18:59,412] Trial 29 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:02,214] Trial 30 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:04,989] Trial 31 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:07,699] Trial 32 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:10,396] Trial 33 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:13,137] Trial 34 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:13,479] Trial 35 finished with value: 0.48239960158212314 and parameters: {'n_es
timators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:16,186] Trial 36 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:16,521] Trial 37 finished with value: 0.48239960158212314 and parameters: {'n_es
timators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:19,292] Trial 38 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:19,635] Trial 39 finished with value: 0.48239960158212314 and parameters: {'n_es
timators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:22,396] Trial 40 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:25,221] Trial 41 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:27,976] Trial 42 finished with value: 0.07511800964286763 and parameters: {'n_es
```

```
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:30,683] Trial 43 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:33,460] Trial 44 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:33,801] Trial 45 finished with value: 0.48239960158212314 and parameters: {'n_es
timators': 10, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:36,459] Trial 46 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:39,221] Trial 47 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:41,913] Trial 48 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
[I 2021-11-10 22:19:44,652] Trial 49 finished with value: 0.07511800964286763 and parameters: {'n_es
timators': 100, 'learning_rate': 0.01}. Best is trial 0 with value: 0.07511800964286763.
```

In [13]:
```
print(f"\tBest value: {study.best_value:.5f}")
print(f"\tBest params:")

for key, value in study.best_params.items():
    print(f"\t\t{key}: {value}")
```

```
        Best value: 0.07512
        Best params:
                n_estimators: 100
                learning_rate: 0.01
```

In [14]:
```
model = lightgbm.LGBMClassifier(objective="multi_logloss", **study.best_params)
model.fit(X_train, y_train)
```

Out[14]:
```
LGBMClassifier(learning_rate=0.01, objective='multi_logloss')
```

In [15]:
```
print(f"Training Accuracy: {model.score(X_train, y_train):0.5f}")
print(f"Test Accuracy: {model.score(X_test, y_test):0.5f}")
```

```
Training Accuracy: 0.95646
Test Accuracy: 0.92164
```

In [ ]:
```

```