DSAA 5002 - Data Mining and Knowledge Discovery in Data Science

Final Exam Report – Q5 Smoke Status Recognition

50015940 Jiaxiang Gao

1. Data Preprocessing:

a) The 'id' column is removed from the training dataset as it's not relevant for prediction.

b) Missing values in various columns are handled:

- For 'height(cm)' and 'waist(cm)', missing values are filled with the mean of the respective columns.
- For 'eyesight(left)' and 'eyesight(right)', missing values in one eye are replaced with values from the other eye.
- For 'hearing(right)', missing values are replaced with values from 'hearing(left)'.
- Missing values in 'Urine protein' are filled with the mean.

c) Duplicate rows in the training dataset are identified and removed.

2. Model Training and Validation:

a) The data is scaled using StandardScaler to normalize feature values.

b) The K-Fold cross-validation approach (with 10 splits) is applied to validate the model's performance.

c) Two types of models are trained and validated:

i. LGBMClassifier: A Light Gradient Boosting Machine classifier.

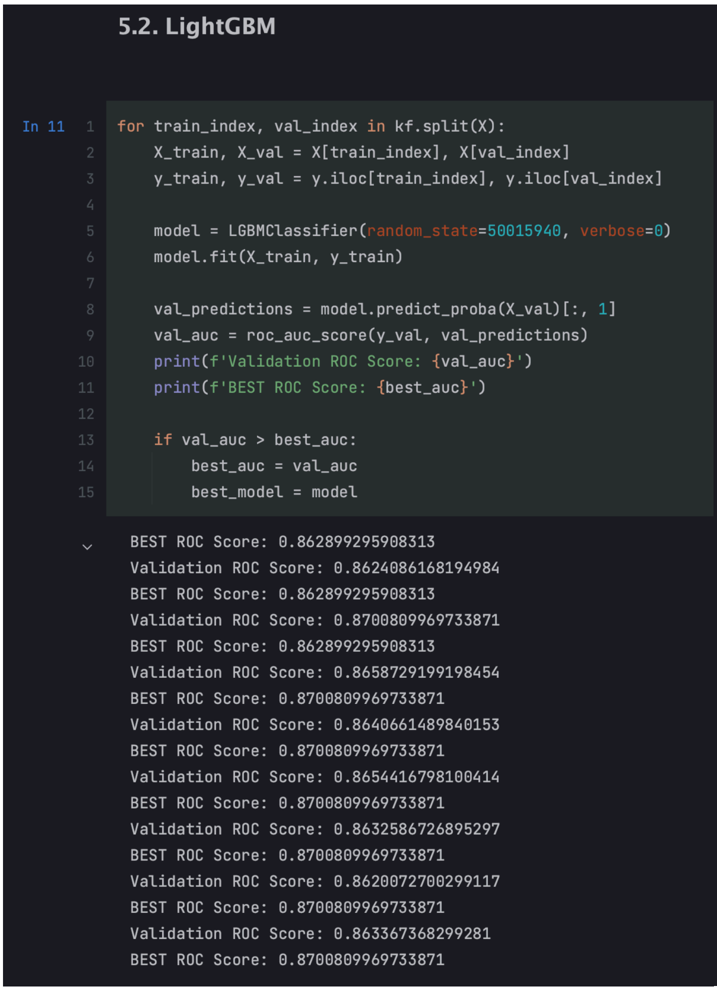
ii. CatBoostClassifier: A classifier from the CatBoost framework.

d) In each fold of the cross-validation:

- The model is trained on the training subset.
- The model's performance is evaluated on the validation subset using the ROC AUC score.
- The best model is updated if the current model's ROC AUC score is higher than the previously recorded best score.

3. Final Model Prediction:

The model with the highest ROC AUC score from the cross-validation process is selected as the best model. This best model is used to predict the 'smoking' variable on the test dataset.



5.2. LightGBM

In 11

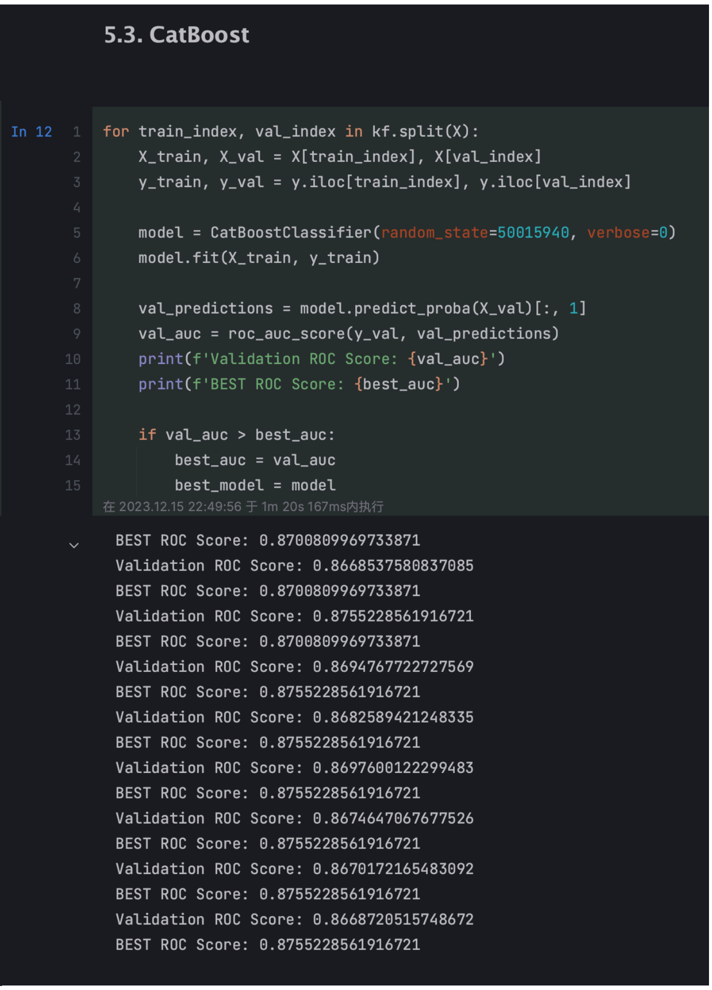
```python
for train_index, val_index in kf.split(X):
    X_train, X_val = X[train_index], X[val_index]
    y_train, y_val = y.iloc[train_index], y.iloc[val_index]

    model = LGBMClassifier(random_state=50015940, verbose=0)
    model.fit(X_train, y_train)

    val_predictions = model.predict_proba(X_val)[:, 1]
    val_auc = roc_auc_score(y_val, val_predictions)
    print(f'Validation ROC Score: {val_auc}')
    print(f'BEST ROC Score: {best_auc}')

    if val_auc > best_auc:
        best_auc = val_auc
        best_model = model
```

```
BEST ROC Score: 0.862899295908313
Validation ROC Score: 0.8624086168194984
BEST ROC Score: 0.862899295908313
Validation ROC Score: 0.8700809969733871
BEST ROC Score: 0.862899295908313
Validation ROC Score: 0.8658729199198454
BEST ROC Score: 0.8700809969733871
Validation ROC Score: 0.8640661489840153
BEST ROC Score: 0.8700809969733871
Validation ROC Score: 0.8654416798100414
BEST ROC Score: 0.8700809969733871
Validation ROC Score: 0.8632586726895297
BEST ROC Score: 0.8700809969733871
Validation ROC Score: 0.8620072700299117
BEST ROC Score: 0.8700809969733871
Validation ROC Score: 0.863367368299281
BEST ROC Score: 0.8700809969733871
```



5.3. CatBoost

In 12

```python
for train_index, val_index in kf.split(X):
    X_train, X_val = X[train_index], X[val_index]
    y_train, y_val = y.iloc[train_index], y.iloc[val_index]

    model = CatBoostClassifier(random_state=50015940, verbose=0)
    model.fit(X_train, y_train)

    val_predictions = model.predict_proba(X_val)[:, 1]
    val_auc = roc_auc_score(y_val, val_predictions)
    print(f'Validation ROC Score: {val_auc}')
    print(f'BEST ROC Score: {best_auc}')

    if val_auc > best_auc:
        best_auc = val_auc
        best_model = model
```

在 2023.12.15 22:49:56 于 1m 20s 167ms内执行

```
BEST ROC Score: 0.8700809969733871
Validation ROC Score: 0.8668537580837085
BEST ROC Score: 0.8700809969733871
Validation ROC Score: 0.8755228561916721
BEST ROC Score: 0.8700809969733871
Validation ROC Score: 0.8694767722727569
BEST ROC Score: 0.8755228561916721
Validation ROC Score: 0.8682589421248335
BEST ROC Score: 0.8755228561916721
Validation ROC Score: 0.8697600122299483
BEST ROC Score: 0.8755228561916721
Validation ROC Score: 0.8674647067677526
BEST ROC Score: 0.8755228561916721
Validation ROC Score: 0.8670172165483092
BEST ROC Score: 0.8755228561916721
Validation ROC Score: 0.8668720515748672
BEST ROC Score: 0.8755228561916721
```