Empowering Financial Insights: Predicting Customer Transactions with Advanced Machine Learning

In this notebook, we aim to predict whether customers will make a specific transaction in the future using machine learning techniques. Our approach involves various data preprocessing techniques and model building strategies to optimize performance. We will explore multiple model versions, each with different strategies for handling the data and feature engineering.

In [2]:
```
import numpy as np
import pandas as pd
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from sklearn.model_selection import StratifiedKFold, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import f1_score, roc_auc_score, roc_curve, accuracy_score, cla
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns
import joblib
```

In [3]:
```
train = pd.read_csv(r"train.csv")
test = pd.read_csv(r"test.csv")
```

In [4]:
```
train.sample(10)
```

Out[4]:

	ID_code	target	var_0	var_1	var_2	var_3	var_4	var_5	var_6
126291	train_126291	1	7.1061	2.5162	9.7902	9.1782	10.1327	-2.1839	4.3226
44068	train_44068	0	14.6245	0.7622	15.6593	7.4749	11.2130	-9.0417	4.8924
65269	train_65269	0	17.5887	-4.2776	12.3604	4.5615	12.8307	-4.7195	5.4608
94036	train_94036	0	9.0451	-5.3304	12.3644	6.8955	11.3978	-3.8517	5.4273
62832	train_62832	0	14.0789	-6.6696	12.0120	6.4697	10.8467	8.5718	5.9157
41362	train_41362	1	10.3317	-0.3173	15.2826	8.4846	13.4219	-16.5936	5.8558
36245	train_36245	1	11.1535	2.5683	14.3768	3.5758	10.8850	-2.2093	6.3655
184062	train_184062	0	15.5129	-4.1668	9.6569	5.8974	11.0991	-12.0629	4.0331
6260	train_6260	0	9.8054	-3.5607	9.0193	7.8273	11.7178	0.8777	6.5975
70774	train_70774	0	10.7270	1.2081	8.5852	7.7302	9.3600	7.9731	4.0607

10 rows × 202 columns

In [5]:
```
# Analysis
train.shape, test.shape
```

Out[5]: ((200000, 202), (200000, 201))

In [6]:
```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Columns: 202 entries, ID_code to var_199
dtypes: float64(200), int64(1), object(1)
memory usage: 308.2+ MB
```

In [7]:
```
train.describe()
```

Out[7]:

	target	var_0	var_1	var_2	var_3	
count	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000	200000.
mean	0.100490	10.679914	-1.627622	10.715192	6.796529	11.
std	0.300653	3.040051	4.050044	2.640894	2.043319	1.
min	0.000000	0.408400	-15.043400	2.117100	-0.040200	5.
25%	0.000000	8.453850	-4.740025	8.722475	5.254075	9.
50%	0.000000	10.524750	-1.608050	10.580000	6.825000	11.
75%	0.000000	12.758200	1.358625	12.516700	8.324100	12.
max	1.000000	20.315000	10.376800	19.353000	13.188300	16.

8 rows × 201 columns

Data Preprocessing: Optimizing Features for Better Predictions

In order to prepare the data for modeling, we must clean and optimize it. Our data preprocessing steps include:

1. **Dropping the ID column:** The `ID_code` column is not useful for prediction.
2. **Handling missing values:** We fill any missing values with the mean of the respective feature to maintain data integrity.
3. **Removing highly correlated features:** To reduce multicollinearity, we eliminate features with a correlation higher than 0.9. This ensures that redundant features do not negatively affect model performance.

The `clean_train()` function is responsible for these steps.

In [9]:
```
# check duplicates
train.duplicated().sum()
```

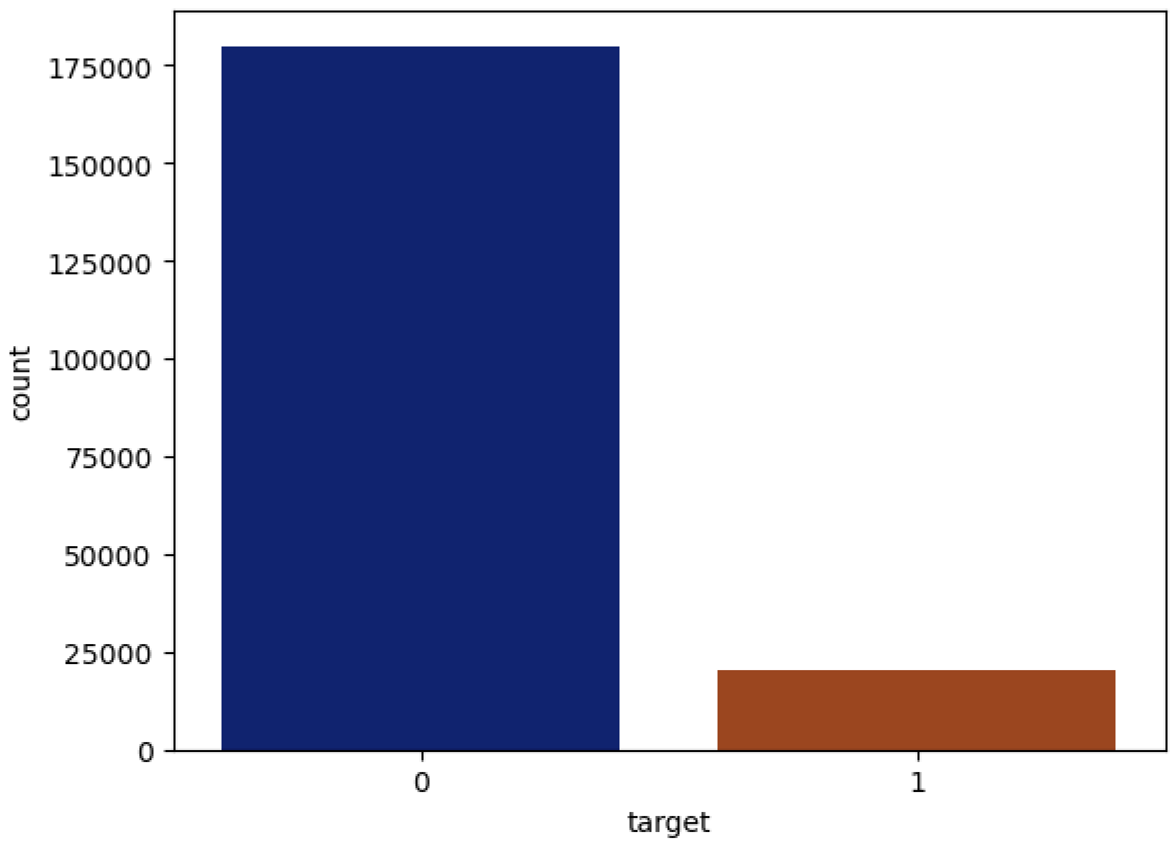
Out[9]: 0

In [10]:
```python
# check missing values
train.isnull().sum().values
```

Out[10]:
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0], dtype=int64)
```

In [11]:
```python
# check imbalance
sns.countplot(train,x='target', palette='dark')
```

Out[11]:
```
<Axes: xlabel='target', ylabel='count'>
```



In [12]:
```python
# Clean and preprocess the train
def clean_train(df, corr_threshold=0.9):
    # Drop ID column (not useful for prediction)
    df = df.drop(columns=['ID_code'])

    # Check for null values and fill if necessary
    df.fillna(df.mean(), inplace=True)

    # Drop highly correlated features (correlation threshold > 0.9)
    corr_matrix = df.corr().abs()
```

```
    upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(bool)
    to_drop = [column for column in upper.columns if any(upper[column] > corr_thres

    df = df.drop(columns=to_drop)

    return df
```

In [13]:
```
# Clean train
df = clean_train(train)
```

In [14]:
```
X = df.drop(columns=['target']).values
y = df['target']
```

Version 1: Baseline Model Without Fake Rows

In this version, we build a baseline model using only the original training data without any artificial (fake) rows added. This serves as a control to understand the performance of a simple model before introducing more sophisticated techniques.

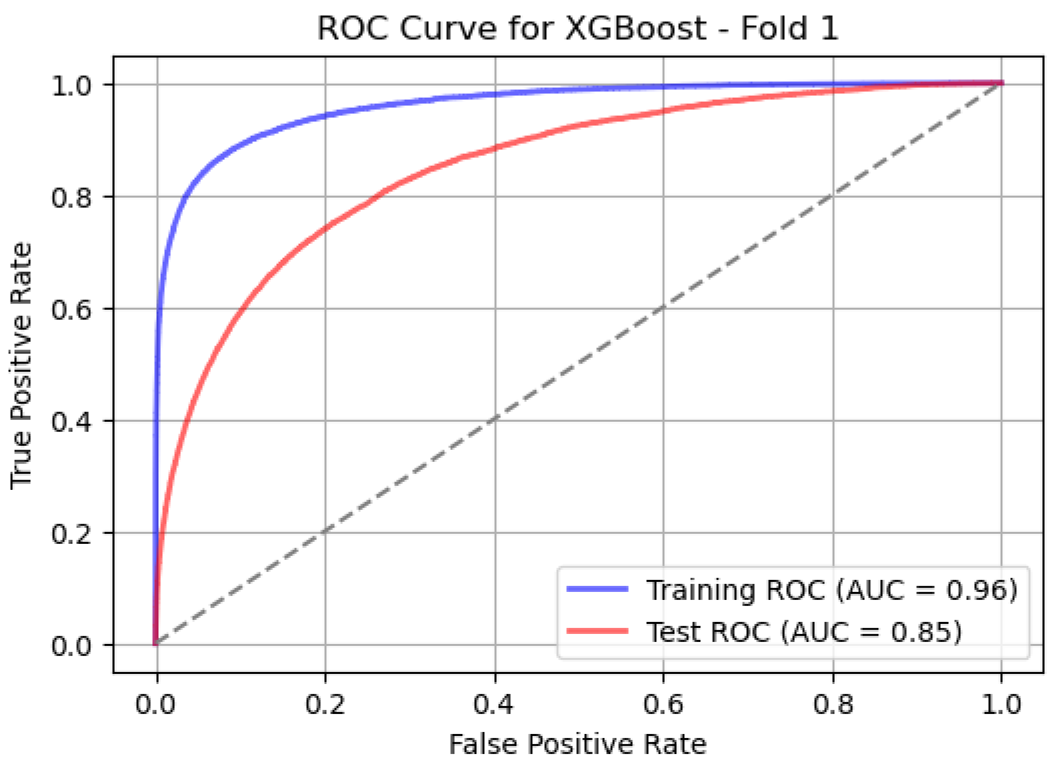
- We clean the training data using the `clean_train()` function.
- The baseline model is built using XGBoost and LightGBM, two popular gradient boosting algorithms.
- Performance metrics such as AUC-ROC and F1 score are used to evaluate the model.

In [16]:
```
def train_and_evaluate_models_v1(X, y, n_splits=2):
    skf = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=42)
    models = {
        'XGBoost': XGBClassifier(n_estimators=100, learning_rate=0.1, max_depth=6,
        'LightGBM': LGBMClassifier(learning_rate=0.04, num_leaves=31, max_bin=1023,
    }

    fold_metrics = {name: {'accuracy': [], 'f1_score': [], 'roc_auc': []} for name
    mean_fpr = np.linspace(0, 1, 100)
    tpr_list_train = {name: [] for name in models.keys()}
    tpr_list_test = {name: [] for name in models.keys()}

    fold_number = 1
    for train_index, test_index in skf.split(X, y):
        X_train, X_test = X[train_index], X[test_index]
        y_train, y_test = y.iloc[train_index], y.iloc[test_index]

        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train)
        X_test_scaled = scaler.transform(X_test)

        for name, model in models.items():
            model.fit(X_train_scaled, y_train)
            y_train_pred = model.predict( X_train_scaled)
            y_train_pred_prob = model.predict_proba( X_train_scaled)[:, 1]
            train_accuracy = accuracy_score(y_train, y_train_pred)
            train_f1 = f1_score(y_train, y_train_pred)
            train_auc = roc_auc_score(y_train, y_train_pred_prob)
```

```
            y_test_pred = model.predict(X_test_scaled)
            y_test_pred_prob = model.predict_proba(X_test_scaled)[:, 1]
            test_accuracy = accuracy_score(y_test, y_test_pred)
            test_f1 = f1_score(y_test, y_test_pred)
            test_auc = roc_auc_score(y_test, y_test_pred_prob)

            fold_metrics[name]['accuracy'].append(test_accuracy)
            fold_metrics[name]['f1_score'].append(test_f1)
            fold_metrics[name]['roc_auc'].append(test_auc)

            fpr_train, tpr_train, _ = roc_curve(y_train, y_train_pred_prob)
            fpr_test, tpr_test, _ = roc_curve(y_test, y_test_pred_prob)
            tpr_list_train[name].append(np.interp(mean_fpr, fpr_train, tpr_train))
            tpr_list_train[name][-1][0] = 0.0
            tpr_list_test[name].append(np.interp(mean_fpr, fpr_test, tpr_test))
            tpr_list_test[name][-1][0] = 0.0

            plt.figure(figsize=(6, 4))
            plt.plot(fpr_train, tpr_train, color='blue', label=f'Training ROC (AUC
            plt.plot(fpr_test, tpr_test, color='red', label=f'Test ROC (AUC = {test
            plt.plot([0, 1], [0, 1], color='gray', linestyle='--')
            plt.xlabel('False Positive Rate')
            plt.ylabel('True Positive Rate')
            plt.title(f'ROC Curve for {name} - Fold {fold_number}')
            plt.legend(loc='lower right')
            plt.grid()
            plt.show()

            print(f"{name} Fold {fold_number} Metrics:")
            print(f"Training Accuracy: {train_accuracy:.2f}, Test Accuracy: {test_a
            print(f"Training F1 Score: {train_f1:.2f}, Test F1 Score: {test_f1:.2f}
            print(f"Training AUC: {train_auc:.2f}, Test AUC: {test_auc:.2f}")
            print(f"Classification Report for Test Set:\n{classification_report(y_t
            if abs(train_auc - test_auc) > 0.10:
                print("Warning: Possible Overfitting Detected")
        fold_number += 1

    plt.figure(figsize=(6, 4))
    for name in models.keys():
        mean_tpr_train = np.mean(tpr_list_train[name], axis=0)
        mean_tpr_train[-1] = 1.0
        mean_auc_train = auc(mean_fpr, mean_tpr_train)
        plt.plot(mean_fpr, mean_tpr_train, lw=2, linestyle='-', label=f'{name} Mean

        mean_tpr_test = np.mean(tpr_list_test[name], axis=0)
        mean_tpr_test[-1] = 1.0
        mean_auc_test = auc(mean_fpr, mean_tpr_test)
        plt.plot(mean_fpr, mean_tpr_test, lw=2, linestyle='-', label=f'{name} Mean

    plt.plot([0, 1], [0, 1], color='gray', linestyle='--')
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('Final Mean ROC Curve for All Models')
    plt.legend(loc='lower right')
    plt.grid()
```

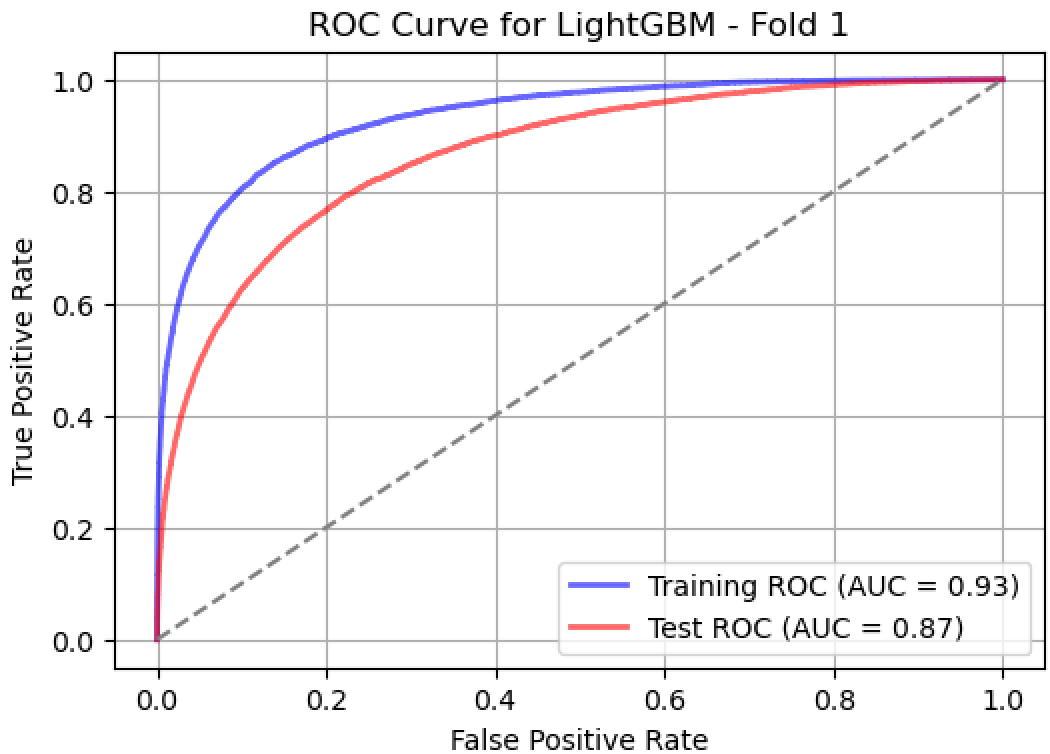
```
    plt.show()

    for name, metrics in fold_metrics.items():
        final_accuracy = np.mean(metrics['accuracy'])
        final_f1 = np.mean(metrics['f1_score'])
        final_roc_auc = np.mean(metrics['roc_auc'])
        print(f"\n{name} Final Cross-Validation Metrics:")
        print(f"Final Accuracy: {final_accuracy:.2f}")
        print(f"Final F1 Score: {final_f1:.2f}")
        print(f"Final ROC AUC: {final_roc_auc:.2f}")

    return models, scaler
```

In [17]:
```
models, scaler = train_and_evaluate_models_v1(X, y)
```

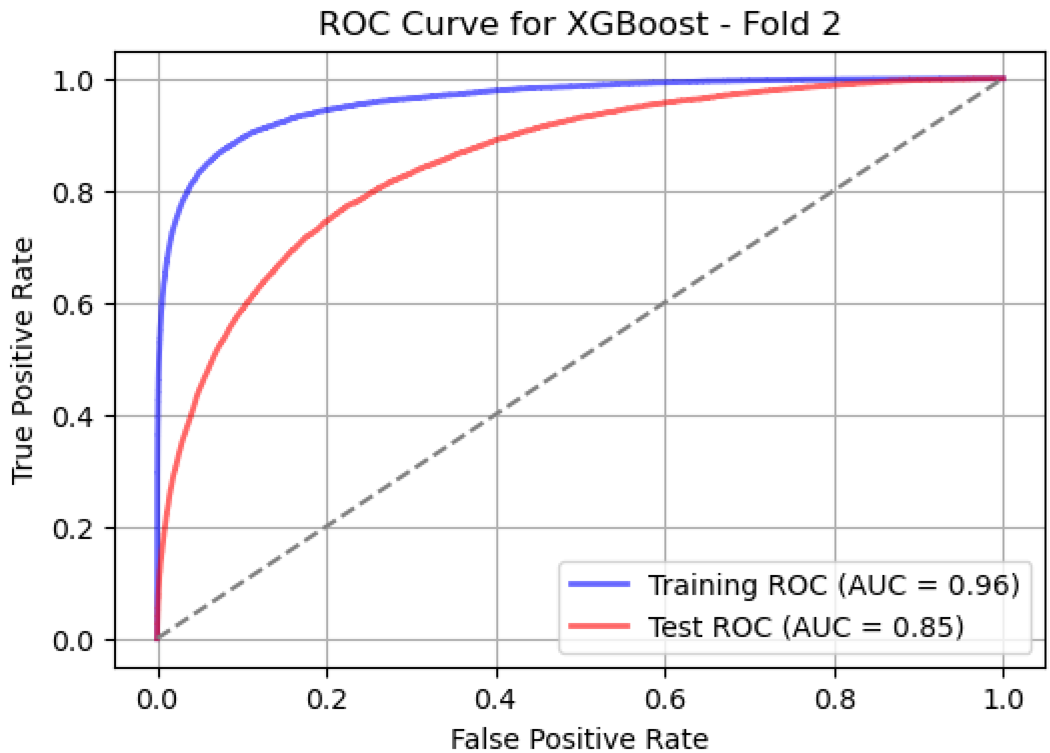


```
XGBoost Fold 1 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.91
Training F1 Score: 0.45, Test F1 Score: 0.17
Training AUC: 0.96, Test AUC: 0.85
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.91      1.00      0.95     89951
           1       0.85      0.10      0.17     10049

    accuracy                           0.91    100000
   macro avg       0.88      0.55      0.56    100000
weighted avg       0.90      0.91      0.87    100000

Warning: Possible Overfitting Detected
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 10049, number of negative: 89951
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.281139 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 203987
[LightGBM] [Info] Number of data points in the train set: 100000, number of used fea
tures: 200
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.100490 -> initscore=-2.191792
[LightGBM] [Info] Start training from score -2.191792
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
```

```
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

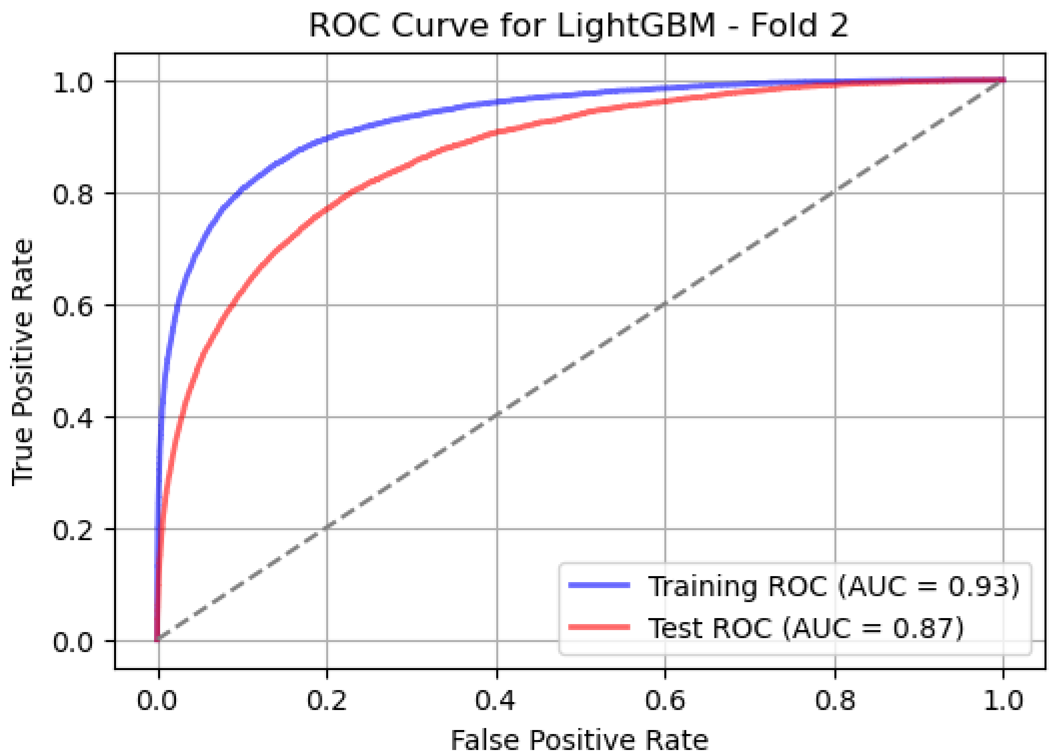



```
LightGBM Fold 1 Metrics:
Training Accuracy: 0.91, Test Accuracy: 0.90
Training F1 Score: 0.22, Test F1 Score: 0.11
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.90      1.00      0.95     89951
           1       0.92      0.06      0.11     10049

    accuracy                           0.90    100000
   macro avg       0.91      0.53      0.53    100000
weighted avg       0.91      0.90      0.87    100000
```

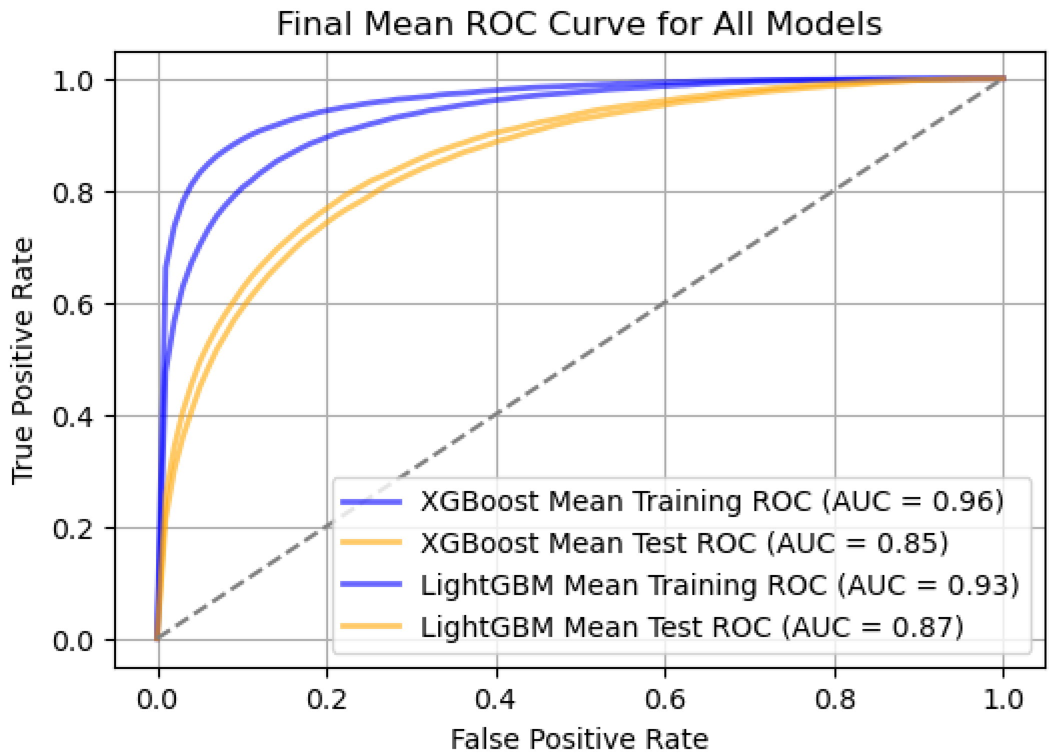

ROC Curve for XGBoost - Fold 2
True Positive Rate
False Positive Rate
Training ROC (AUC = 0.96)
Test ROC (AUC = 0.85)
0.0
0.2
0.4
0.6
0.8
1.0

```
XGBoost Fold 2 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.91
Training F1 Score: 0.44, Test F1 Score: 0.17
Training AUC: 0.96, Test AUC: 0.85
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.91      1.00      0.95     89951
           1       0.84      0.09      0.17     10049

    accuracy                           0.91    100000
   macro avg       0.87      0.54      0.56    100000
weighted avg       0.90      0.91      0.87    100000

Warning: Possible Overfitting Detected
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 10049, number of negative: 89951
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.378510 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 203985
[LightGBM] [Info] Number of data points in the train set: 100000, number of used fea
tures: 200
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.100490 -> initscore=-2.191792
[LightGBM] [Info] Start training from score -2.191792
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
```

```
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```




```
LightGBM Fold 2 Metrics:
Training Accuracy: 0.91, Test Accuracy: 0.90
Training F1 Score: 0.22, Test F1 Score: 0.11
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.90      1.00      0.95     89951
           1       0.90      0.06      0.11     10049

    accuracy                           0.90    100000
   macro avg       0.90      0.53      0.53    100000
weighted avg       0.90      0.90      0.86    100000
```




```
XGBoost Final Cross-Validation Metrics:
Final Accuracy: 0.91
Final F1 Score: 0.17
Final ROC AUC: 0.85

LightGBM Final Cross-Validation Metrics:
Final Accuracy: 0.90
Final F1 Score: 0.11
Final ROC AUC: 0.87
```

In [ ]:

Version 2: Model with Cleaned Dataset (Without PCA)

In this version, we focus on cleaning the dataset by removing duplicate rows and outliers before training the model. We do **not** apply Principal Component Analysis (PCA) in this version, opting to retain the full feature set.

- **Data Cleaning:** We ensure the integrity of the dataset by checking for and removing duplicate rows and outliers (using z-scores). This helps improve the model's performance by eliminating noisy or abnormal data points that could negatively affect learning.
- **Model Training:** We use the same gradient boosting algorithms as before, but now train the models on this cleaned dataset. The goal is to assess how the model performs without dimensionality reduction through PCA and with a more refined dataset.

In [19]:
```
# Check for fake rows based on duplicated rows or abnormal values
def check_fake_rows(df):
    # Check for duplicate rows
```

```
    duplicate_rows = df[df.duplicated()]
    print(f"Number of duplicate rows: {len(duplicate_rows)}")
    if len(duplicate_rows) > 0:
        print("Duplicate rows found:\n", duplicate_rows)
        df = df.drop_duplicates()

    # Check for outliers in the train using z-score
    from scipy.stats import zscore
    z_scores = np.abs(df.apply(zscore))
    df = df[(z_scores < 3).all(axis=1)]
    outliers = (z_scores > 3).sum().sum()
    print(f"Number of potential outliers (removed) (z-score > 3): {outliers}")

    return df
```

In [20]:

```
# Check for fake rows
df1 = check_fake_rows(df)
```

```
Number of duplicate rows: 0
Number of potential outliers (removed) (z-score > 3): 11299
```

In [21]:

```
# Prepare features and target
X_1 = df1.drop(columns=['target']).values
y_1 = df1['target']
```

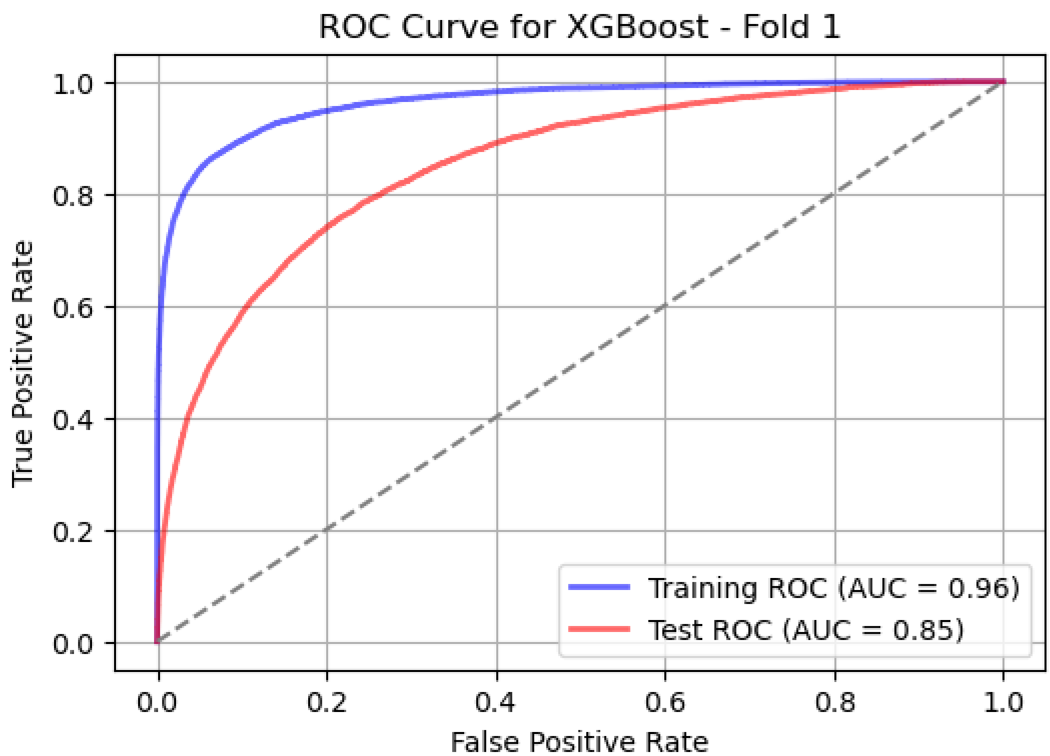
In [22]:

```
def train_and_evaluate_models_v2(X, y, n_splits=2):
    skf = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=42)
    models = {
        'XGBoost': XGBClassifier(n_estimators=100, learning_rate=0.1, max_depth=6,
        'LightGBM': LGBMClassifier(learning_rate=0.04, num_leaves=31, max_bin=1023,
    }

    fold_metrics = {name: {'accuracy': [], 'f1_score': [], 'roc_auc': []} for name
    mean_fpr = np.linspace(0, 1, 100)
    tpr_list_train = {name: [] for name in models.keys()}
    tpr_list_test = {name: [] for name in models.keys()}

    fold_number = 1
    for train_index, test_index in skf.split(X, y):
        X_train, X_test = X[train_index], X[test_index]
        y_train, y_test = y.iloc[train_index], y.iloc[test_index]

        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train)
        X_test_scaled = scaler.transform(X_test)

        for name, model in models.items():
            model.fit(X_train_scaled, y_train)
            y_train_pred = model.predict( X_train_scaled)
            y_train_pred_prob = model.predict_proba( X_train_scaled)[:, 1]
            train_accuracy = accuracy_score(y_train, y_train_pred)
            train_f1 = f1_score(y_train, y_train_pred)
            train_auc = roc_auc_score(y_train, y_train_pred_prob)

            y_test_pred = model.predict(X_test_scaled)
            y_test_pred_prob = model.predict_proba(X_test_scaled)[:, 1]
```

```
        test_accuracy = accuracy_score(y_test, y_test_pred)
        test_f1 = f1_score(y_test, y_test_pred)
        test_auc = roc_auc_score(y_test, y_test_pred_prob)

        fold_metrics[name]['accuracy'].append(test_accuracy)
        fold_metrics[name]['f1_score'].append(test_f1)
        fold_metrics[name]['roc_auc'].append(test_auc)

        fpr_train, tpr_train, _ = roc_curve(y_train, y_train_pred_prob)
        fpr_test, tpr_test, _ = roc_curve(y_test, y_test_pred_prob)
        tpr_list_train[name].append(np.interp(mean_fpr, fpr_train, tpr_train))
        tpr_list_train[name][-1][0] = 0.0
        tpr_list_test[name].append(np.interp(mean_fpr, fpr_test, tpr_test))
        tpr_list_test[name][-1][0] = 0.0

        plt.figure(figsize=(6, 4))
        plt.plot(fpr_train, tpr_train, color='blue', label=f'Training ROC (AUC
        plt.plot(fpr_test, tpr_test, color='red', label=f'Test ROC (AUC = {test
        plt.plot([0, 1], [0, 1], color='gray', linestyle='--')
        plt.xlabel('False Positive Rate')
        plt.ylabel('True Positive Rate')
        plt.title(f'ROC Curve for {name} - Fold {fold_number}')
        plt.legend(loc='lower right')
        plt.grid()
        plt.show()

        print(f"{name} Fold {fold_number} Metrics:")
        print(f"Training Accuracy: {train_accuracy:.2f}, Test Accuracy: {test_a
        print(f"Training F1 Score: {train_f1:.2f}, Test F1 Score: {test_f1:.2f}
        print(f"Training AUC: {train_auc:.2f}, Test AUC: {test_auc:.2f}")
        print(f"Classification Report for Test Set:\n{classification_report(y_t
        if abs(train_auc - test_auc) > 0.10:
            print("Warning: Possible Overfitting Detected")
    fold_number += 1

plt.figure(figsize=(6, 4))
for name in models.keys():
    mean_tpr_train = np.mean(tpr_list_train[name], axis=0)
    mean_tpr_train[-1] = 1.0
    mean_auc_train = auc(mean_fpr, mean_tpr_train)
    plt.plot(mean_fpr, mean_tpr_train, lw=2, linestyle='-', label=f'{name} Mean

    mean_tpr_test = np.mean(tpr_list_test[name], axis=0)
    mean_tpr_test[-1] = 1.0
    mean_auc_test = auc(mean_fpr, mean_tpr_test)
    plt.plot(mean_fpr, mean_tpr_test, lw=2, linestyle='-', label=f'{name} Mean

plt.plot([0, 1], [0, 1], color='gray', linestyle='--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Final Mean ROC Curve for All Models')
plt.legend(loc='lower right')
plt.grid()
plt.show()

for name, metrics in fold_metrics.items():
```

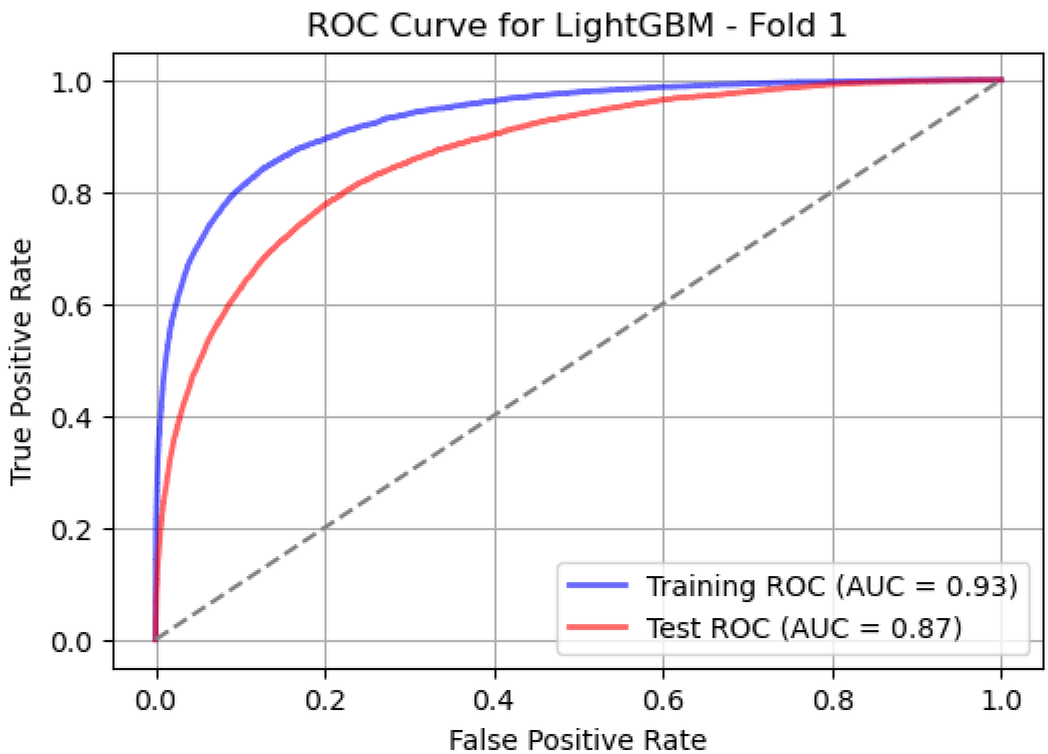
```
        final_accuracy = np.mean(metrics['accuracy'])
        final_f1 = np.mean(metrics['f1_score'])
        final_roc_auc = np.mean(metrics['roc_auc'])
        print(f"\n{name} Final Cross-Validation Metrics:")
        print(f"Final Accuracy: {final_accuracy:.2f}")
        print(f"Final F1 Score: {final_f1:.2f}")
        print(f"Final ROC AUC: {final_roc_auc:.2f}")

    return models
```

In [23]:
```
train_and_evaluate_models_v2(X_1, y_1)
```

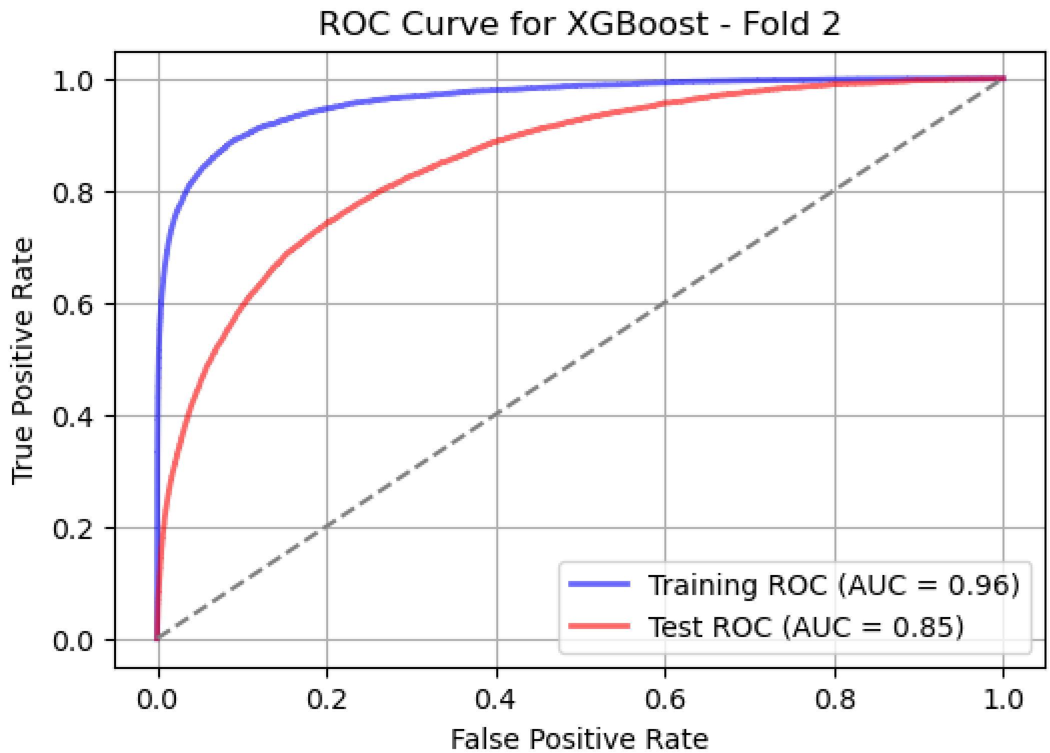


```
XGBoost Fold 1 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.91
Training F1 Score: 0.45, Test F1 Score: 0.16
Training AUC: 0.96, Test AUC: 0.85
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.91      1.00      0.95     85072
           1       0.84      0.09      0.16      9413

    accuracy                           0.91     94485
   macro avg       0.88      0.54      0.56     94485
weighted avg       0.90      0.91      0.87     94485

Warning: Possible Overfitting Detected
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 9413, number of negative: 85071
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.280382 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 203980
[LightGBM] [Info] Number of data points in the train set: 94484, number of used feat
ures: 200
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.099625 -> initscore=-2.201394
[LightGBM] [Info] Start training from score -2.201394
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
```

```
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

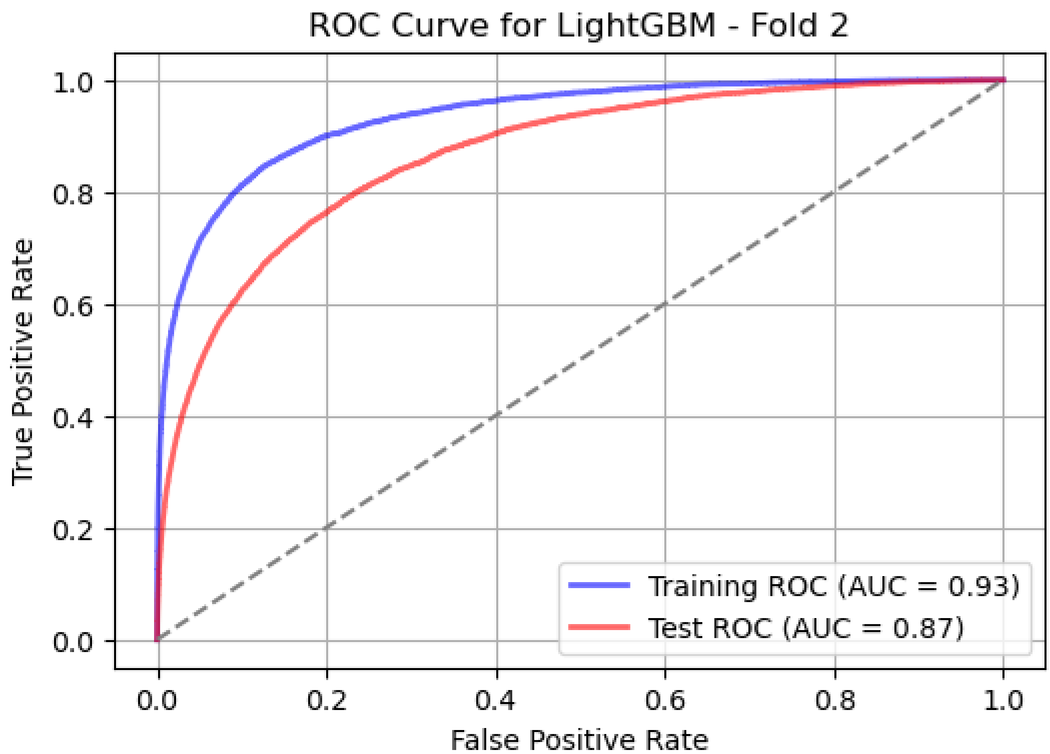



```
LightGBM Fold 1 Metrics:
Training Accuracy: 0.91, Test Accuracy: 0.91
Training F1 Score: 0.23, Test F1 Score: 0.10
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.91      1.00      0.95     85072
           1       0.90      0.05      0.10      9413

    accuracy                           0.91     94485
   macro avg       0.90      0.53      0.53     94485
weighted avg       0.90      0.91      0.87     94485
```

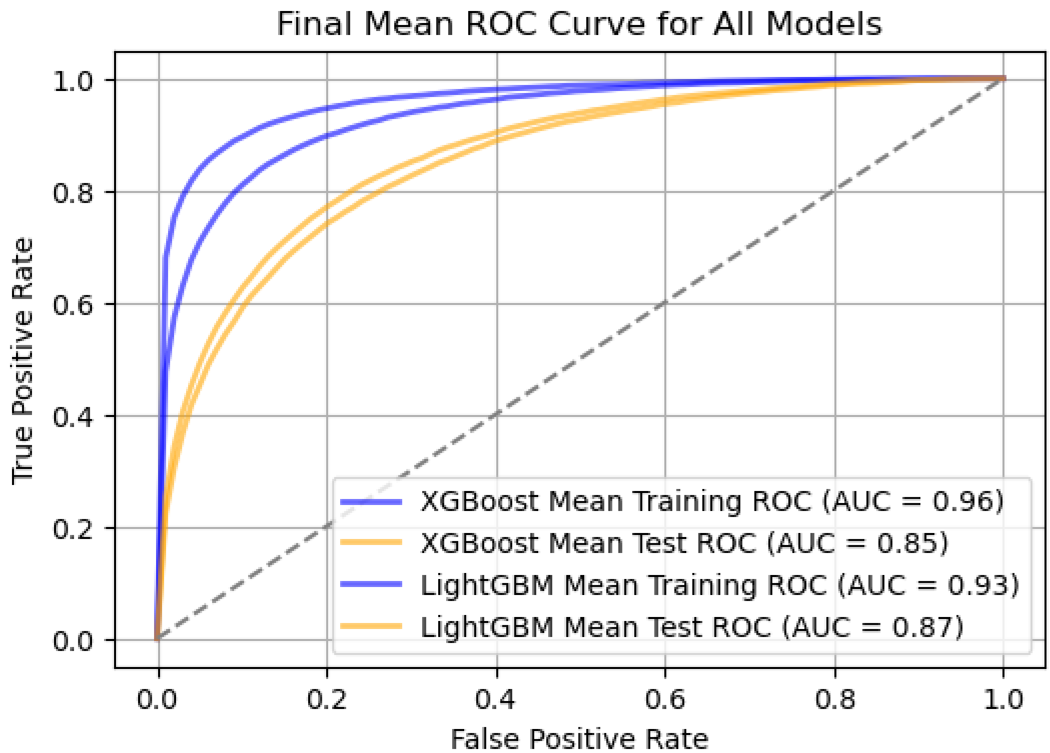

ROC Curve for XGBoost - Fold 2
True Positive Rate
False Positive Rate
Training ROC (AUC = 0.96)
Test ROC (AUC = 0.85)
0.0
0.2
0.4
0.6
0.8
1.0

```
XGBoost Fold 2 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.91
Training F1 Score: 0.46, Test F1 Score: 0.16
Training AUC: 0.96, Test AUC: 0.85
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.91      1.00      0.95     85071
           1       0.83      0.09      0.16      9413

    accuracy                           0.91     94484
   macro avg       0.87      0.54      0.55     94484
weighted avg       0.90      0.91      0.87     94484

Warning: Possible Overfitting Detected
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 9413, number of negative: 85072
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.280734 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 203979
[LightGBM] [Info] Number of data points in the train set: 94485, number of used feat
ures: 200
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.099624 -> initscore=-2.201406
[LightGBM] [Info] Start training from score -2.201406
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
```

```
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```




```
LightGBM Fold 2 Metrics:
Training Accuracy: 0.91, Test Accuracy: 0.91
Training F1 Score: 0.22, Test F1 Score: 0.11
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.91      1.00      0.95     85071
           1       0.90      0.06      0.11      9413

    accuracy                           0.91     94484
   macro avg       0.90      0.53      0.53     94484
weighted avg       0.91      0.91      0.87     94484
```




```
XGBoost Final Cross-Validation Metrics:
Final Accuracy: 0.91
Final F1 Score: 0.16
Final ROC AUC: 0.85

LightGBM Final Cross-Validation Metrics:
Final Accuracy: 0.91
Final F1 Score: 0.10
Final ROC AUC: 0.87
```

Out[23]:

```
{'XGBoost': XGBClassifier(base_score=None, booster=None, callbacks=None,
               colsample_bylevel=None, colsample_bynode=None,
               colsample_bytree=0.8, device=None, early_stopping_rounds=None,
               enable_categorical=False, eval_metric='logloss',
               feature_types=None, gamma=None, grow_policy=None,
               importance_type=None, interaction_constraints=None,
               learning_rate=0.1, max_bin=None, max_cat_threshold=None,
               max_cat_to_onehot=None, max_delta_step=None, max_depth=6,
               max_leaves=None, min_child_weight=None, missing=nan,
               monotone_constraints=None, multi_strategy=None, n_estimators=100,
               n_jobs=None, num_parallel_tree=None, random_state=42, ...),
 'LightGBM': LGBMClassifier(bagging_fraction=0.85, bagging_freq=1, feature_fractio
n=1.0,
                learning_rate=0.04, max_bin=1023, min_child_samples=1000,
                n_estimators=200, n_jobs=-1, objective='binary', reg_alpha=0.1,
                reg_lambda=0.2)}
```

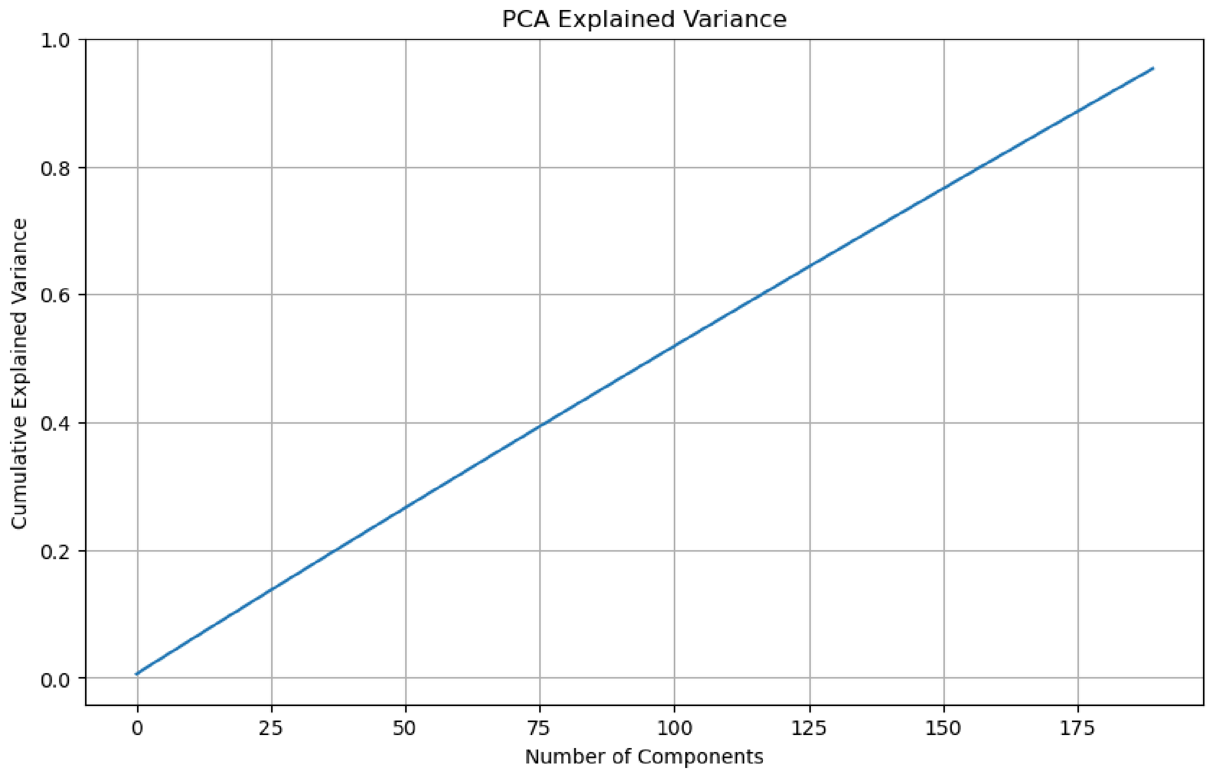
PCA Visualization

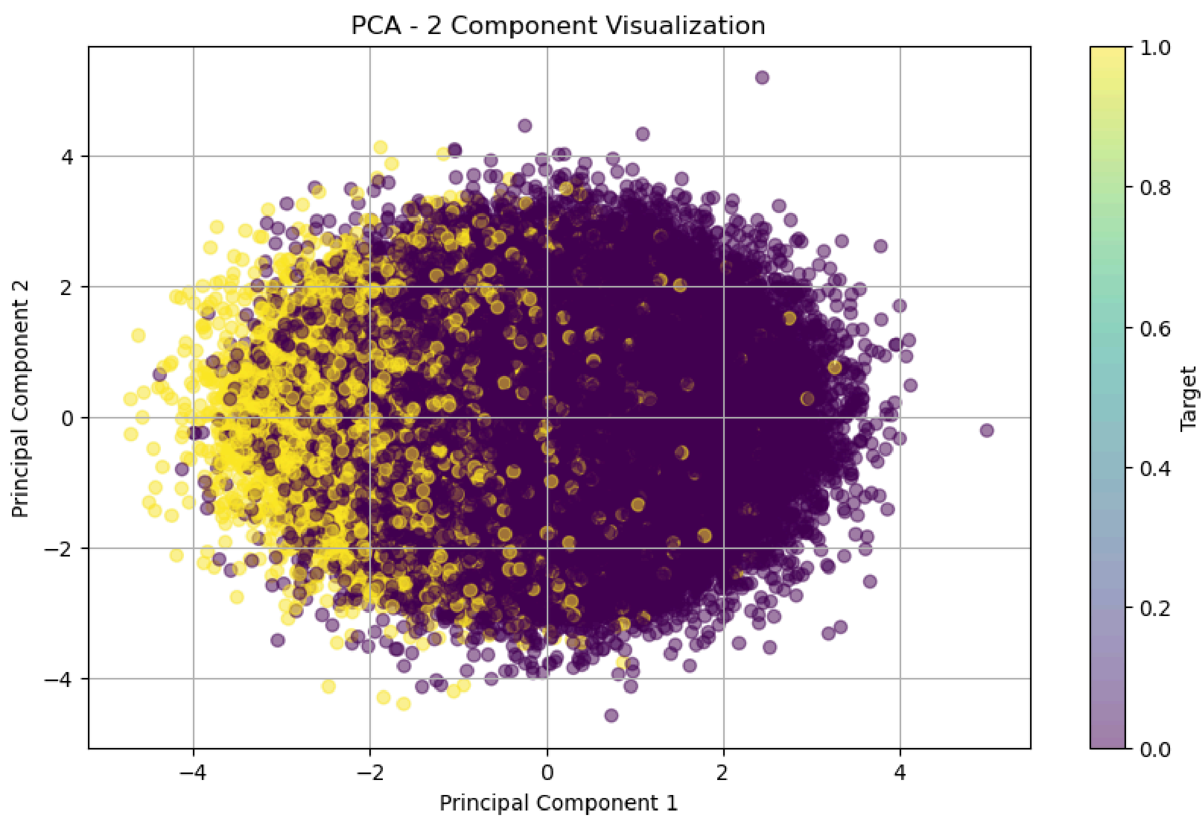
This section presents the PCA (Principal Component Analysis) visualization to better understand the data.

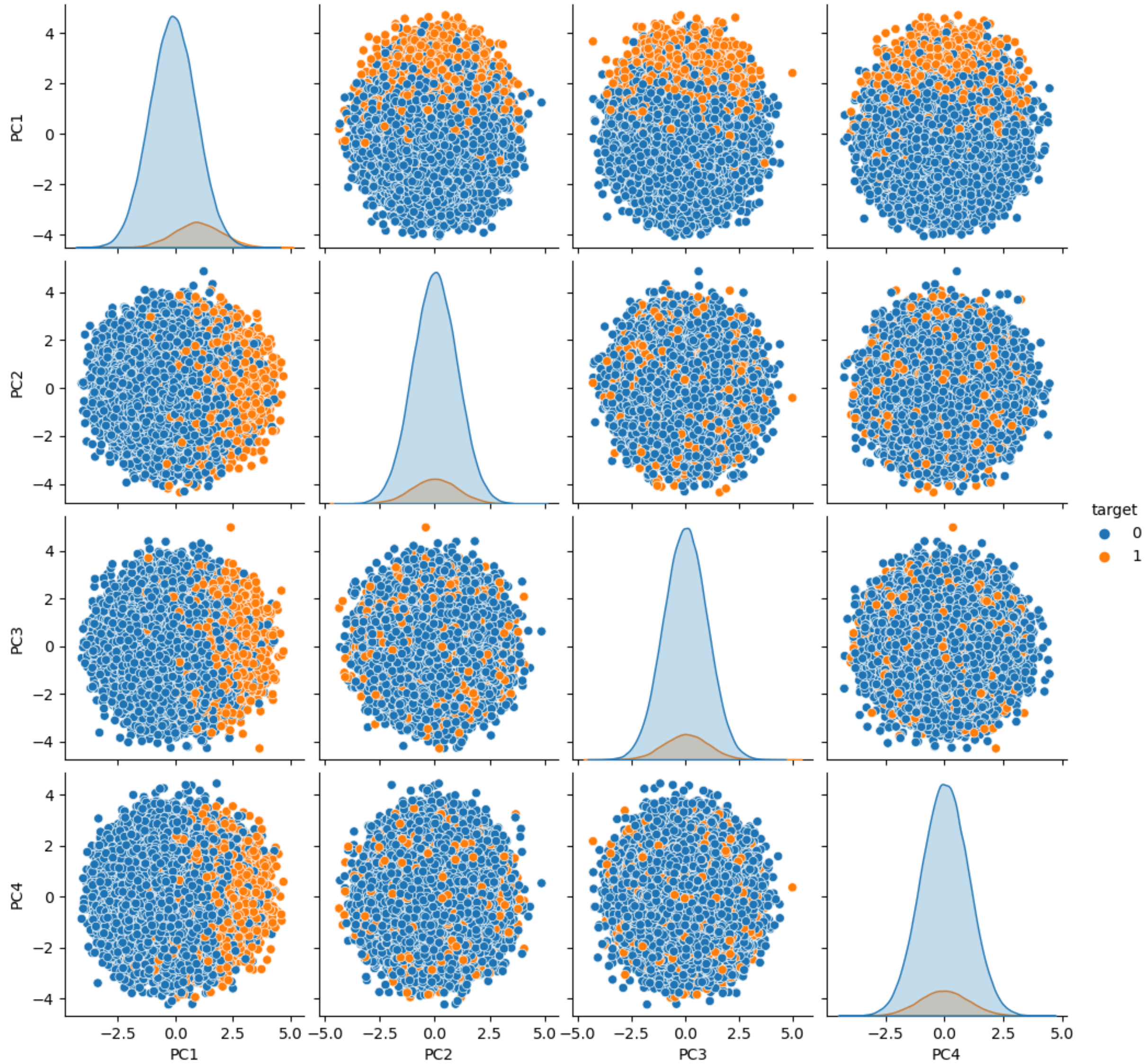
In [26]:
```
# Apply PCA for dimensionality reduction
def apply_pca(X_train, X_test, n_components=0.90):
    pca = PCA(n_components=n_components)
    X_train_pca = pca.fit_transform(X_train)
    X_test_pca = pca.transform(X_test)
    return X_train_pca, X_test_pca, pca
```

In [27]:
```
def visualize_pca_results(X, y):
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)

    # Fit PCA to explain 95% of the variance
    pca = PCA(n_components=0.95)
    X_pca = pca.fit_transform(X_scaled)

    # Explained variance plot
    plt.figure(figsize=(10, 6))
    plt.plot(np.cumsum(pca.explained_variance_ratio_))
    plt.xlabel('Number of Components')
    plt.ylabel('Cumulative Explained Variance')
    plt.title('PCA Explained Variance')
    plt.grid()
    plt.show()

    # 2D visualization of the first two principal components
    pca_2d = PCA(n_components=2)
    X_pca_2d = pca_2d.fit_transform(X_scaled)
    pca_df = pd.DataFrame(X_pca_2d, columns=['PC1', 'PC2'])
    pca_df['target'] = y.values

    plt.figure(figsize=(10, 6))
    plt.scatter(pca_df['PC1'], pca_df['PC2'], c=pca_df['target'], cmap='viridis', a
    plt.xlabel('Principal Component 1')
    plt.ylabel('Principal Component 2')
    plt.title('PCA - 2 Component Visualization')
    plt.colorbar(label='Target')
    plt.grid()
    plt.show()

    # Pairplot for the first four components
    pca_4d = PCA(n_components=4)
    X_pca_4d = pca_4d.fit_transform(X_scaled)
    pca_df_4d = pd.DataFrame(X_pca_4d, columns=['PC1', 'PC2', 'PC3', 'PC4'])
    pca_df_4d['target'] = y.values

    sns.pairplot(pca_df_4d, hue='target', diag_kind='kde')
    plt.show()
```

In [28]:
```
visualize_pca_results(X, y)
```


PCA Explained Variance
Cumulative Explained Variance
Number of Components
0.0
0.2
0.4
0.6
0.8
1.0
0
25
50
75
100
125
150
175


PCA - 2 Component Visualization
Principal Component 2
Principal Component 1
Target
−4
−2
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0



Version 3: Model with PCA

This version builds upon the previous one by applying **Principal Component Analysis (PCA)** to the dataset. PCA is used to reduce the dimensionality of the data while retaining most of the variance, which can improve both the model's training time and its performance by focusing on the most important features.

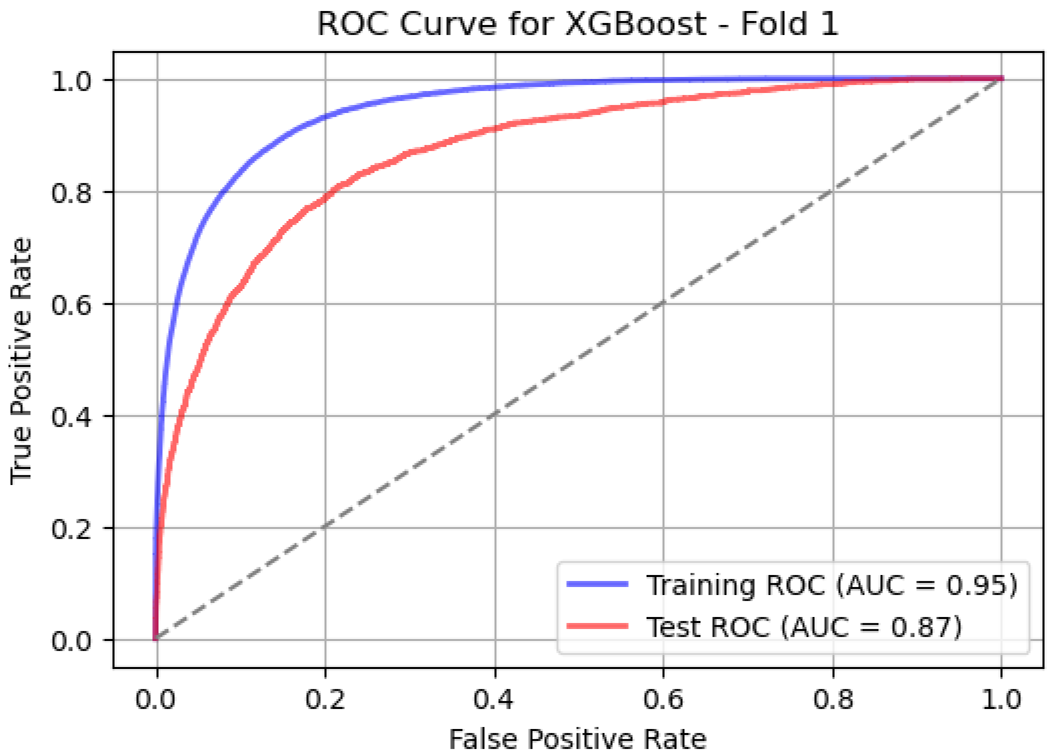
- **PCA Application:** we apply PCA to reduce the feature space while preserving the key variance in the dataset.
- **Model building:** The model is then trained using the transformed dataset (with reduced dimensions) to further improve performance.

In [30]:
```
# Model training function using StratifiedKFold
def train_and_evaluate_models_with_pca(X, y, n_splits=10):
    skf = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=43)
    models = {
        'XGBoost': XGBClassifier(n_estimators=300, learning_rate=0.06, max_depth=6,
```

```
        'LightGBM': LGBMClassifier(learning_rate=0.06, num_leaves=31, max_bin=1023,
    }

    fold_metrics = {name: {'accuracy': [], 'f1_score': [], 'roc_auc': []} for name
    mean_fpr = np.linspace(0, 1, 100)
    tpr_list_train = {name: [] for name in models.keys()}
    tpr_list_test = {name: [] for name in models.keys()}

    fold_number = 1
    for train_index, test_index in skf.split(X, y):
        X_train, X_test = X[train_index], X[test_index]
        y_train, y_test = y.iloc[train_index], y.iloc[test_index]

        # Standardize the data
        scaler = StandardScaler()
        X_train_scaled = scaler.fit_transform(X_train)
        X_test_scaled = scaler.transform(X_test)

        # Apply PCA for dimensionality reduction.
        X_train_pca, X_test_pca, pca = apply_pca(X_train_scaled, X_test_scaled)

        for name, model in models.items():
            model.fit(X_train_pca, y_train)
            y_train_pred = model.predict(X_train_pca)
            y_train_pred_prob = model.predict_proba(X_train_pca)[:, 1]
            train_accuracy = accuracy_score(y_train, y_train_pred)
            train_f1 = f1_score(y_train, y_train_pred)
            train_auc = roc_auc_score(y_train, y_train_pred_prob)

            y_test_pred = model.predict(X_test_pca)
            y_test_pred_prob = model.predict_proba(X_test_pca)[:, 1]
            test_accuracy = accuracy_score(y_test, y_test_pred)
            test_f1 = f1_score(y_test, y_test_pred)
            test_auc = roc_auc_score(y_test, y_test_pred_prob)

            fold_metrics[name]['accuracy'].append(test_accuracy)
            fold_metrics[name]['f1_score'].append(test_f1)
            fold_metrics[name]['roc_auc'].append(test_auc)

            fpr_train, tpr_train, _ = roc_curve(y_train, y_train_pred_prob)
            fpr_test, tpr_test, _ = roc_curve(y_test, y_test_pred_prob)
            tpr_list_train[name].append(np.interp(mean_fpr, fpr_train, tpr_train))
            tpr_list_train[name][-1][0] = 0.0
            tpr_list_test[name].append(np.interp(mean_fpr, fpr_test, tpr_test))
            tpr_list_test[name][-1][0] = 0.0

            plt.figure(figsize=(6, 4))
            plt.plot(fpr_train, tpr_train, color='blue', label=f'Training ROC (AUC
            plt.plot(fpr_test, tpr_test, color='red', label=f'Test ROC (AUC = {test
            plt.plot([0, 1], [0, 1], color='gray', linestyle='--')
            plt.xlabel('False Positive Rate')
            plt.ylabel('True Positive Rate')
            plt.title(f'ROC Curve for {name} - Fold {fold_number}')
            plt.legend(loc='lower right')
            plt.grid()
            plt.show()
```

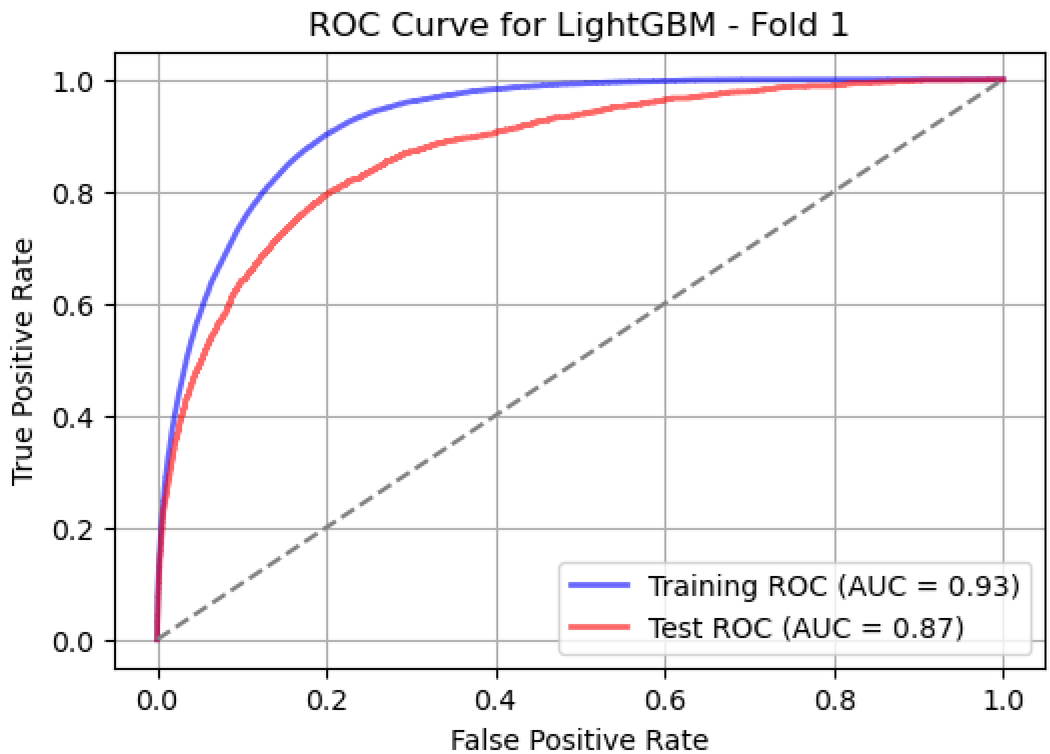
```
            print(f"{name} Fold {fold_number} Metrics:")
            print(f"Training Accuracy: {train_accuracy:.2f}, Test Accuracy: {test_a
            print(f"Training F1 Score: {train_f1:.2f}, Test F1 Score: {test_f1:.2f}
            print(f"Training AUC: {train_auc:.2f}, Test AUC: {test_auc:.2f}")
            print(f"Classification Report for Test Set:\n{classification_report(y_t
            if abs(train_auc - test_auc) > 0.10:
                print("Warning: Possible Overfitting Detected")
        fold_number += 1

    plt.figure(figsize=(6, 4))
    for name in models.keys():
        mean_tpr_train = np.mean(tpr_list_train[name], axis=0)
        mean_tpr_train[-1] = 1.0
        mean_auc_train = auc(mean_fpr, mean_tpr_train)
        plt.plot(mean_fpr, mean_tpr_train, lw=2, linestyle='-', label=f'{name} Mean

        mean_tpr_test = np.mean(tpr_list_test[name], axis=0)
        mean_tpr_test[-1] = 1.0
        mean_auc_test = auc(mean_fpr, mean_tpr_test)
        plt.plot(mean_fpr, mean_tpr_test, lw=2, linestyle='-', label=f'{name} Mean

    plt.plot([0, 1], [0, 1], color='gray', linestyle='--')
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('Final Mean ROC Curve for All Models')
    plt.legend(loc='lower right')
    plt.grid()
    plt.show()

    for name, metrics in fold_metrics.items():
        final_accuracy = np.mean(metrics['accuracy'])
        final_f1 = np.mean(metrics['f1_score'])
        final_roc_auc = np.mean(metrics['roc_auc'])
        print(f"\n{name} Final Cross-Validation Metrics:")
        print(f"Final Accuracy: {final_accuracy:.2f}")
        print(f"Final F1 Score: {final_f1:.2f}")
        print(f"Final ROC AUC: {final_roc_auc:.2f}")

    return models, pca, scaler
```

In [31]:
```
# Train and evaluate models using StratifiedKFold
models, pca, scaler = train_and_evaluate_models_with_pca(X, y)
```

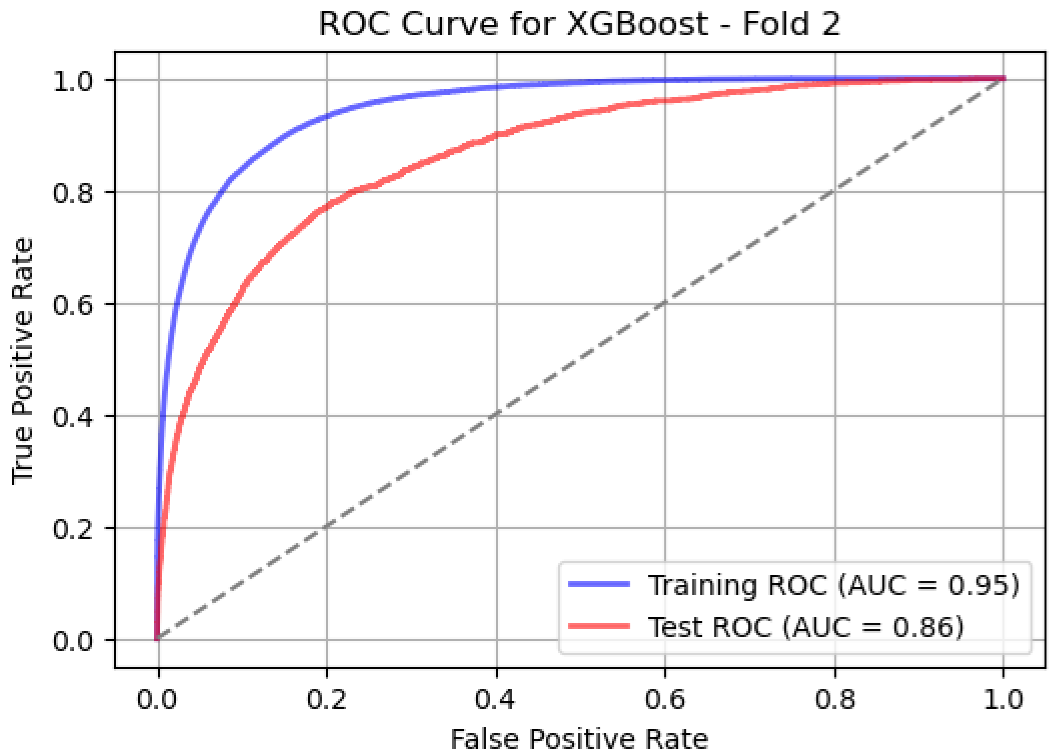

ROC Curve for XGBoost - Fold 1
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Training ROC (AUC = 0.95)
Test ROC (AUC = 0.87)

```
XGBoost Fold 1 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.92
Training F1 Score: 0.54, Test F1 Score: 0.40
Training AUC: 0.95, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     17991
           1       0.69      0.28      0.40      2009

    accuracy                           0.92     20000
   macro avg       0.81      0.63      0.68     20000
weighted avg       0.90      0.92      0.90     20000

[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 18089, number of negative: 161911
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.492916 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 183117
[LightGBM] [Info] Number of data points in the train set: 180000, number of used fea
tures: 179
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=0.000000
[LightGBM] [Info] Start training from score 0.000000
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
```

```
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

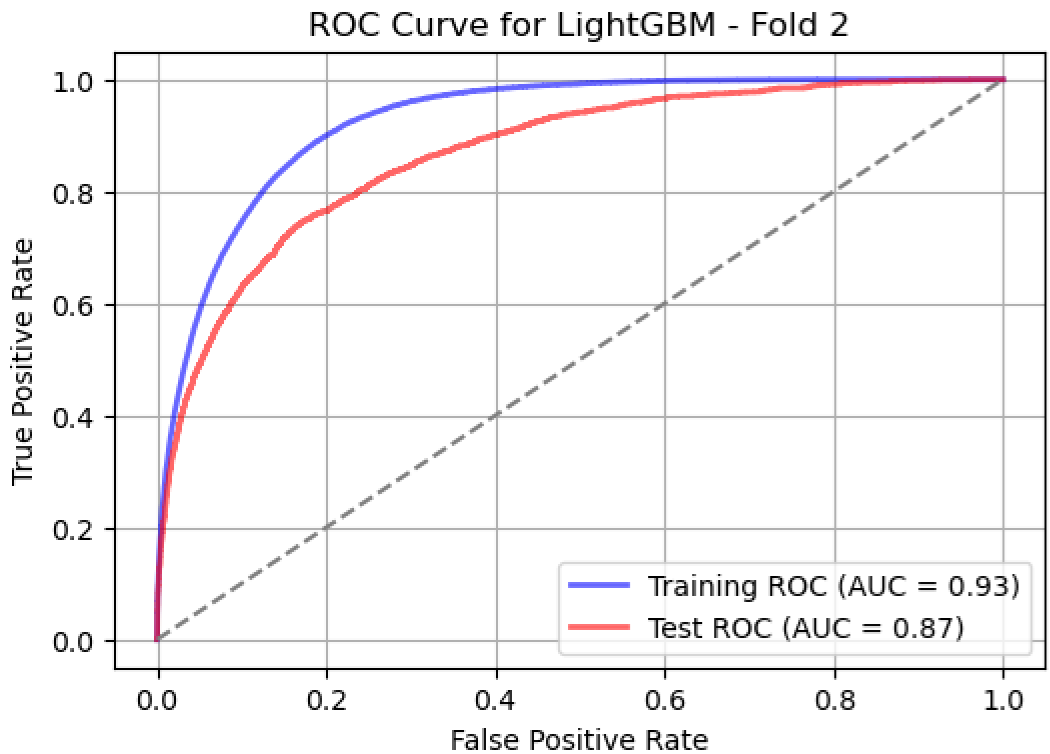



```
LightGBM Fold 1 Metrics:
Training Accuracy: 0.84, Test Accuracy: 0.82
Training F1 Score: 0.52, Test F1 Score: 0.46
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.97      0.83      0.89     17991
           1       0.33      0.76      0.46      2009

    accuracy                           0.82     20000
   macro avg       0.65      0.79      0.68     20000
weighted avg       0.90      0.82      0.85     20000
```

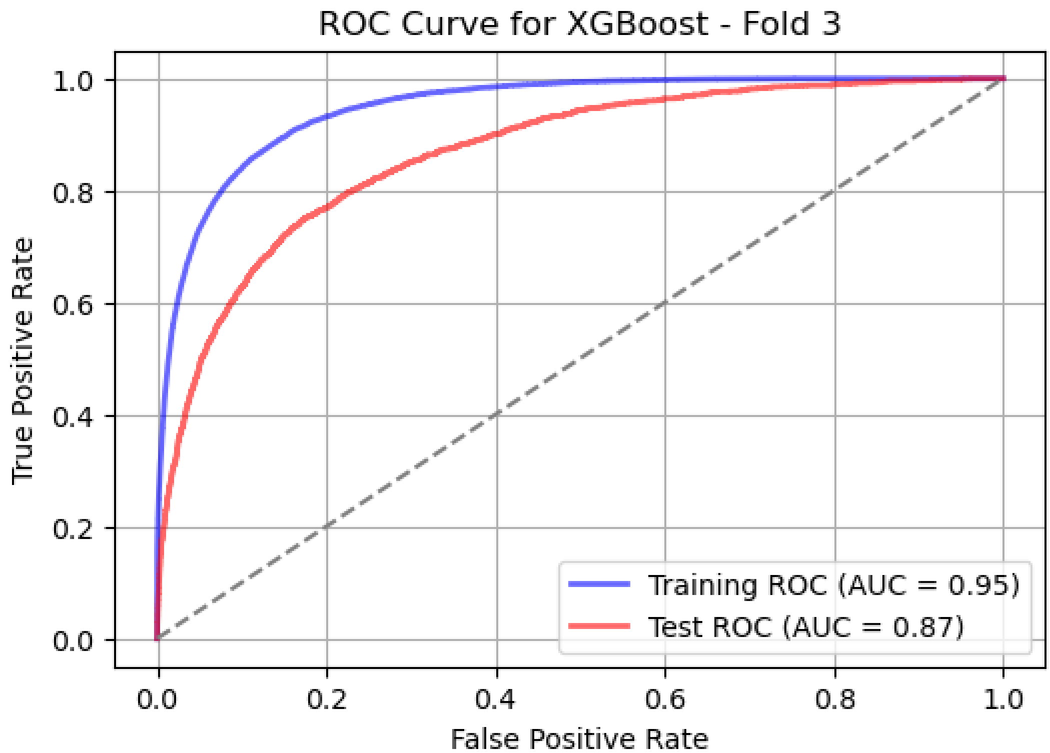

ROC Curve for XGBoost - Fold 2
True Positive Rate
False Positive Rate
Training ROC (AUC = 0.95)
Test ROC (AUC = 0.86)
0.0
0.2
0.4
0.6
0.8
1.0

```
XGBoost Fold 2 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.91
Training F1 Score: 0.54, Test F1 Score: 0.39
Training AUC: 0.95, Test AUC: 0.86
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     17991
           1       0.70      0.27      0.39      2009

    accuracy                           0.91     20000
   macro avg       0.81      0.63      0.67     20000
weighted avg       0.90      0.91      0.90     20000

[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 18089, number of negative: 161911
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.429686 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 183117
[LightGBM] [Info] Number of data points in the train set: 180000, number of used fea
tures: 179
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=0.000000
[LightGBM] [Info] Start training from score 0.000000
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
```

```
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

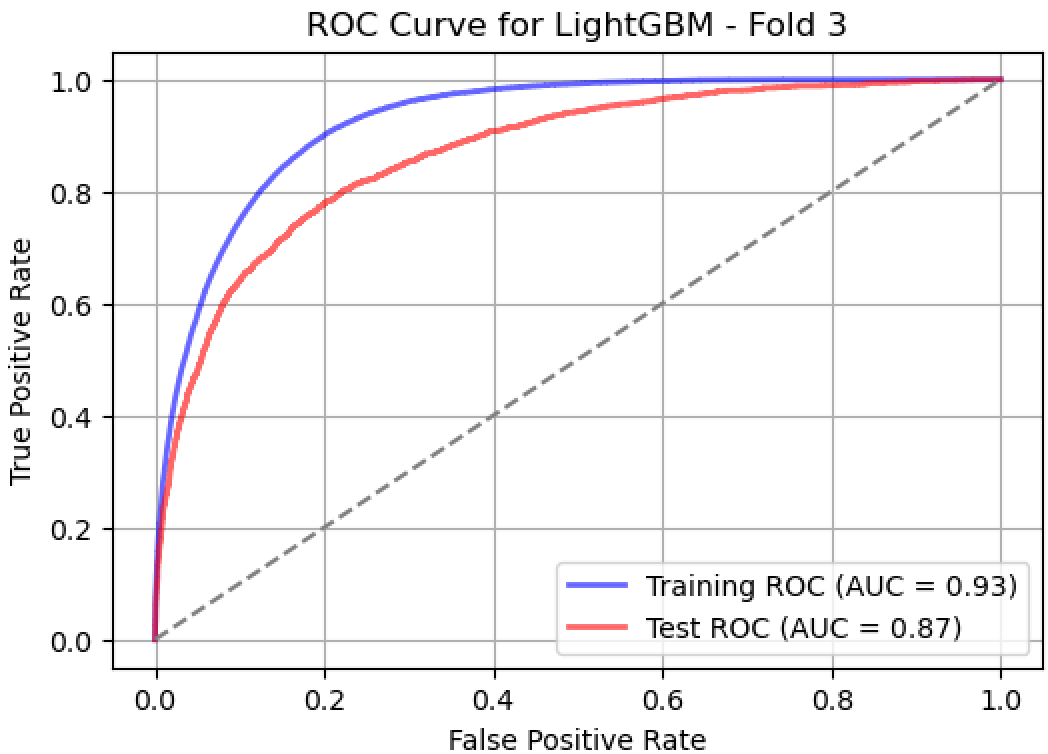



```
LightGBM Fold 2 Metrics:
Training Accuracy: 0.84, Test Accuracy: 0.82
Training F1 Score: 0.52, Test F1 Score: 0.45
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.97      0.83      0.89     17991
           1       0.32      0.75      0.45      2009

    accuracy                           0.82     20000
   macro avg       0.65      0.79      0.67     20000
weighted avg       0.90      0.82      0.85     20000
```

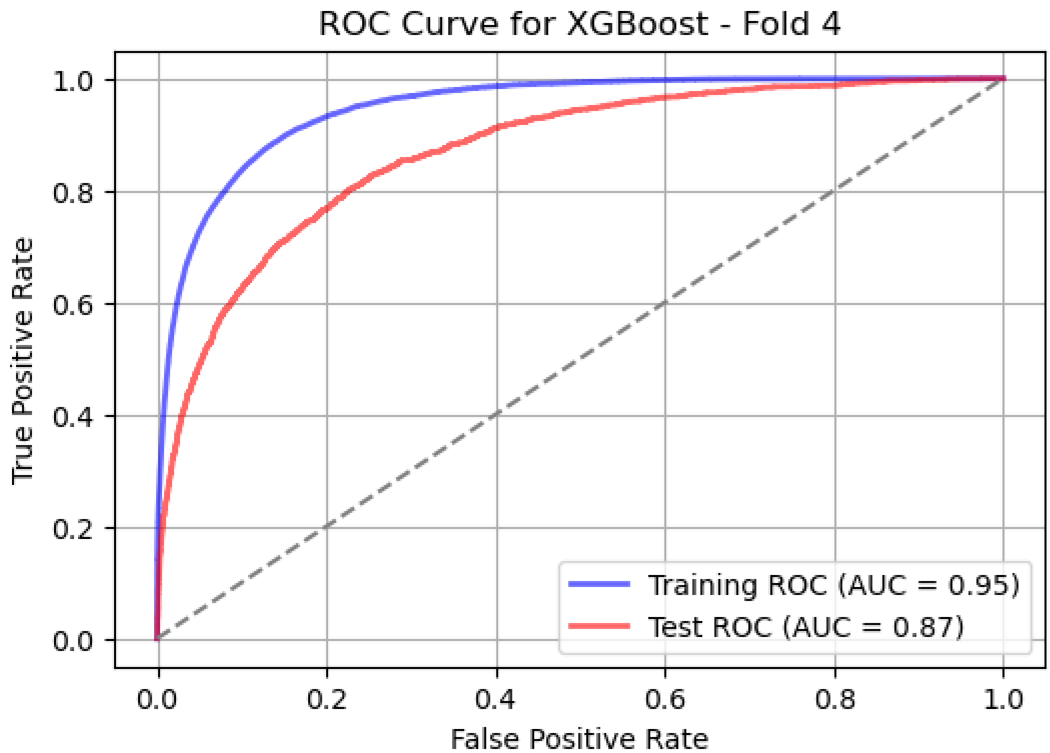

ROC Curve for XGBoost - Fold 3
True Positive Rate
False Positive Rate
Training ROC (AUC = 0.95)
Test ROC (AUC = 0.87)
0.0
0.2
0.4
0.6
0.8
1.0

```
XGBoost Fold 3 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.91
Training F1 Score: 0.54, Test F1 Score: 0.39
Training AUC: 0.95, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.92      0.98      0.95     17990
           1       0.66      0.28      0.39      2010

    accuracy                           0.91     20000
   macro avg       0.79      0.63      0.67     20000
weighted avg       0.90      0.91      0.90     20000

[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 18088, number of negative: 161912
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.428401 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 183117
[LightGBM] [Info] Number of data points in the train set: 180000, number of used fea
tures: 179
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=0.000000
[LightGBM] [Info] Start training from score 0.000000
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
```

```
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

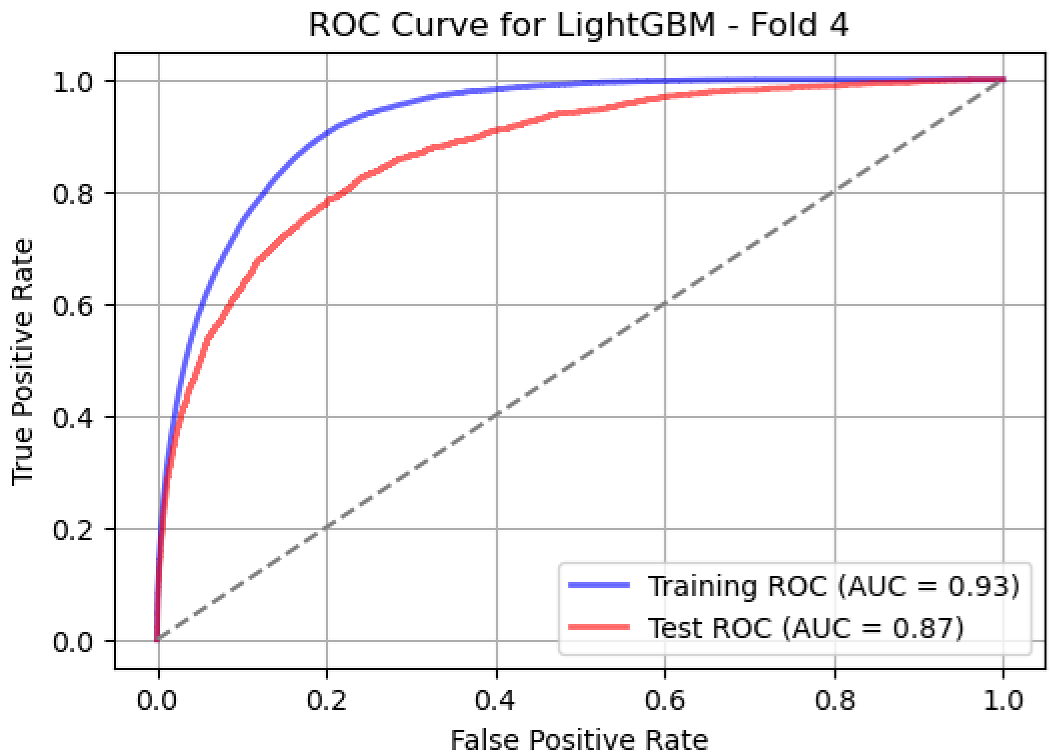



```
LightGBM Fold 3 Metrics:
Training Accuracy: 0.84, Test Accuracy: 0.82
Training F1 Score: 0.52, Test F1 Score: 0.45
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.97      0.82      0.89     17990
           1       0.32      0.75      0.45      2010

    accuracy                           0.82     20000
   macro avg       0.65      0.79      0.67     20000
weighted avg       0.90      0.82      0.85     20000
```

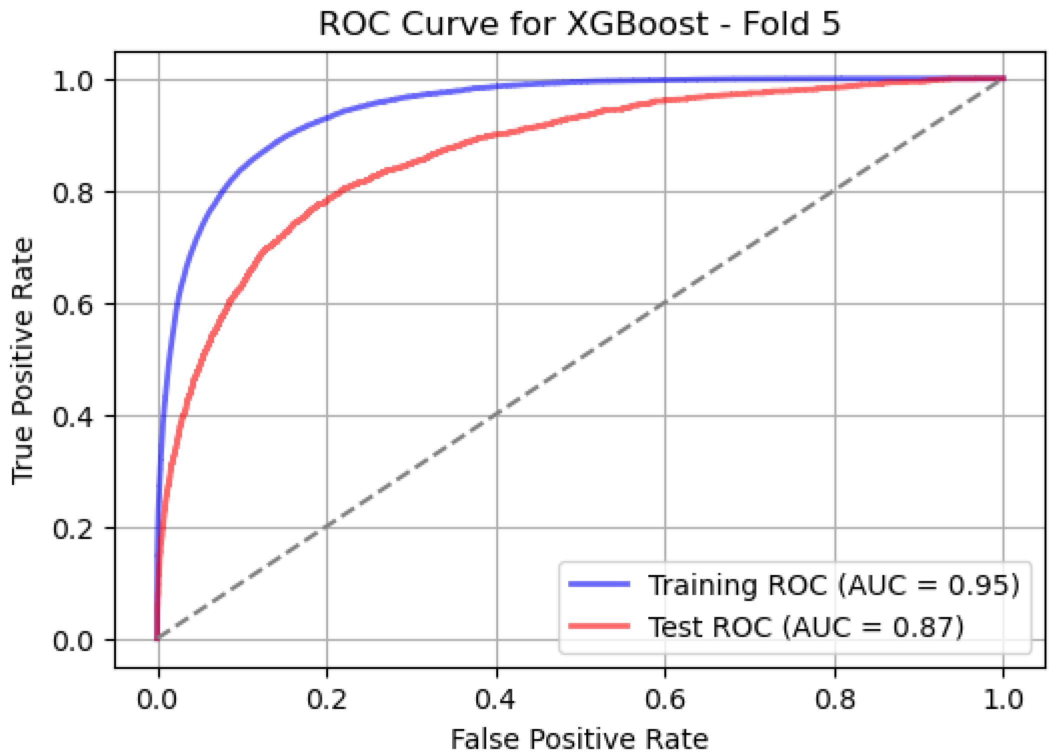

ROC Curve for XGBoost - Fold 4
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Training ROC (AUC = 0.95)
Test ROC (AUC = 0.87)

```
XGBoost Fold 4 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.91
Training F1 Score: 0.54, Test F1 Score: 0.39
Training AUC: 0.95, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     17990
           1       0.69      0.27      0.39      2010

    accuracy                           0.91     20000
   macro avg       0.81      0.63      0.67     20000
weighted avg       0.90      0.91      0.90     20000

[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 18088, number of negative: 161912
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.461017 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 183117
[LightGBM] [Info] Number of data points in the train set: 180000, number of used fea
tures: 179
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=0.000000
[LightGBM] [Info] Start training from score 0.000000
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
```

```
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

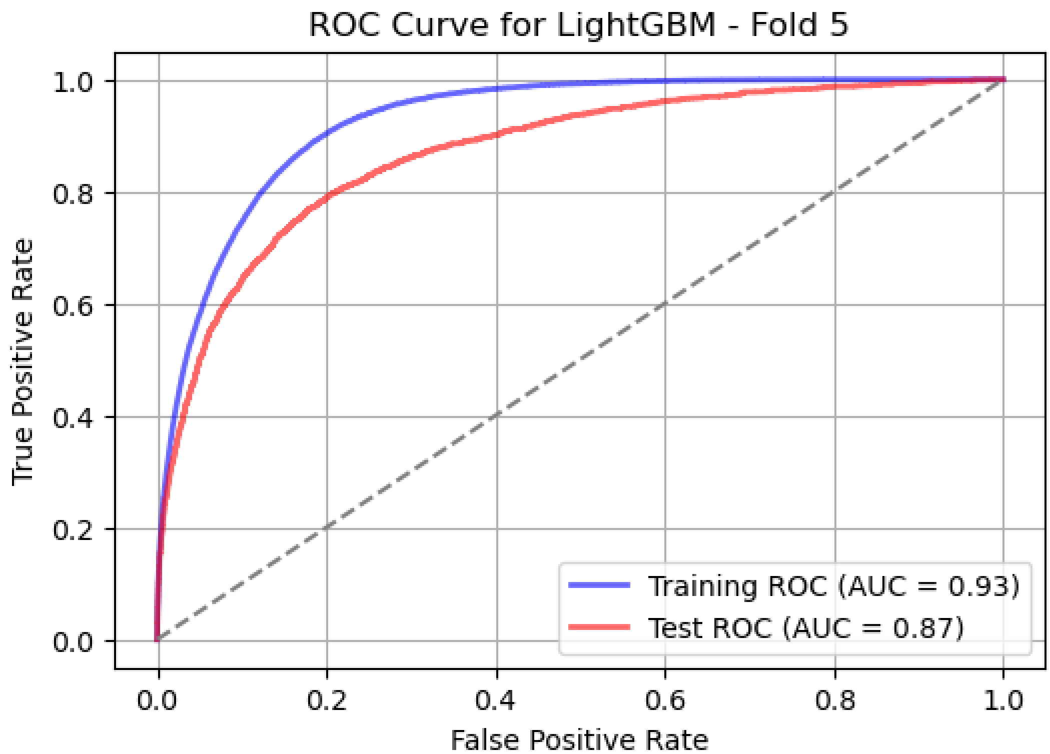



```
LightGBM Fold 4 Metrics:
Training Accuracy: 0.84, Test Accuracy: 0.83
Training F1 Score: 0.52, Test F1 Score: 0.46
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.97      0.84      0.90     17990
           1       0.34      0.73      0.46      2010

    accuracy                           0.83     20000
   macro avg       0.65      0.79      0.68     20000
weighted avg       0.90      0.83      0.85     20000
```

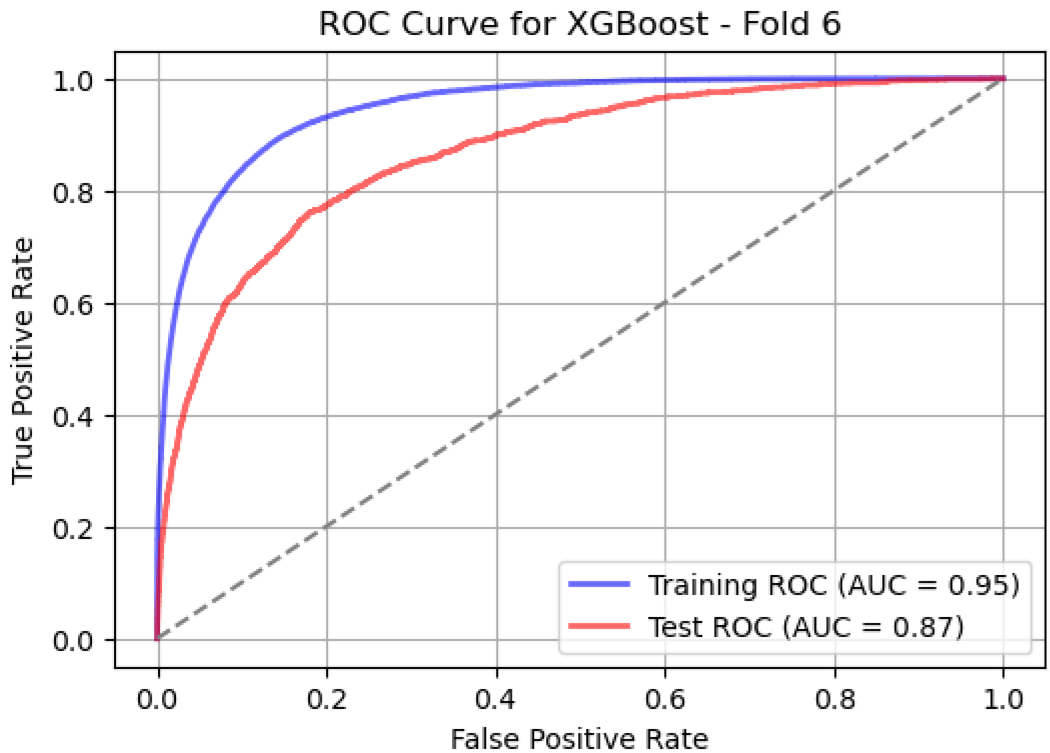

ROC Curve for XGBoost - Fold 5
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Training ROC (AUC = 0.95)
Test ROC (AUC = 0.87)

```
XGBoost Fold 5 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.92
Training F1 Score: 0.54, Test F1 Score: 0.39
Training AUC: 0.95, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     17990
           1       0.70      0.27      0.39      2010

    accuracy                           0.92     20000
   macro avg       0.81      0.63      0.67     20000
weighted avg       0.90      0.92      0.90     20000

[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 18088, number of negative: 161912
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.436246 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 183117
[LightGBM] [Info] Number of data points in the train set: 180000, number of used fea
tures: 179
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=0.000000
[LightGBM] [Info] Start training from score 0.000000
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
```

```
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

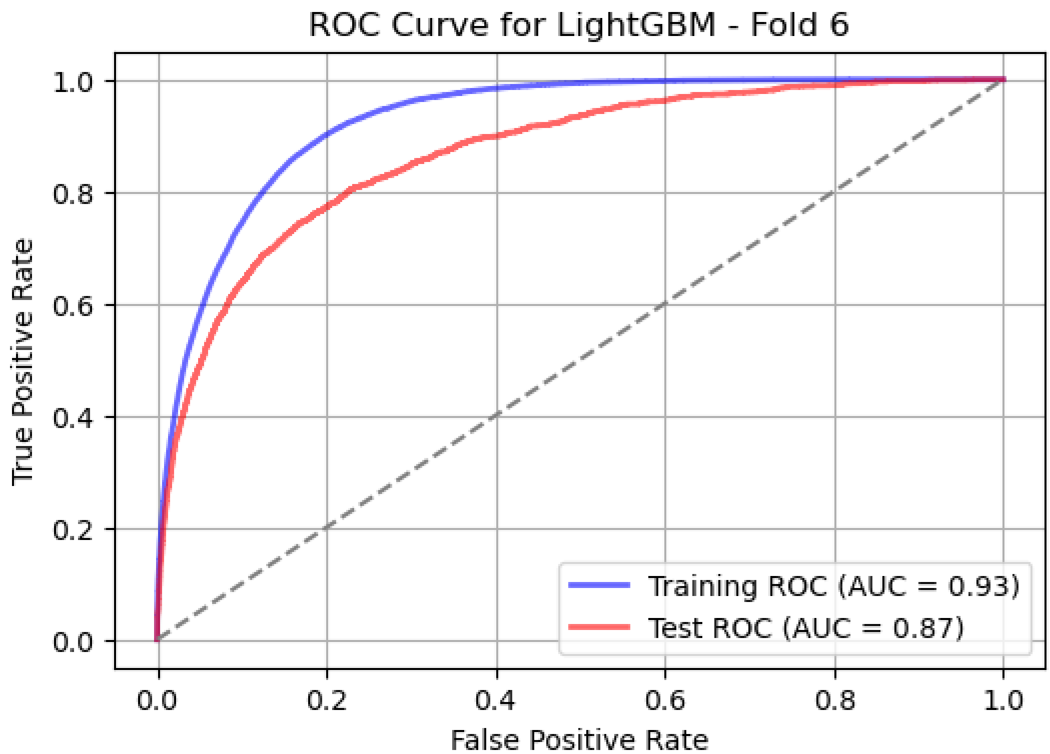



```
LightGBM Fold 5 Metrics:
Training Accuracy: 0.84, Test Accuracy: 0.83
Training F1 Score: 0.52, Test F1 Score: 0.46
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.97      0.84      0.90     17990
           1       0.34      0.75      0.46      2010

    accuracy                           0.83     20000
   macro avg       0.65      0.79      0.68     20000
weighted avg       0.90      0.83      0.85     20000
```

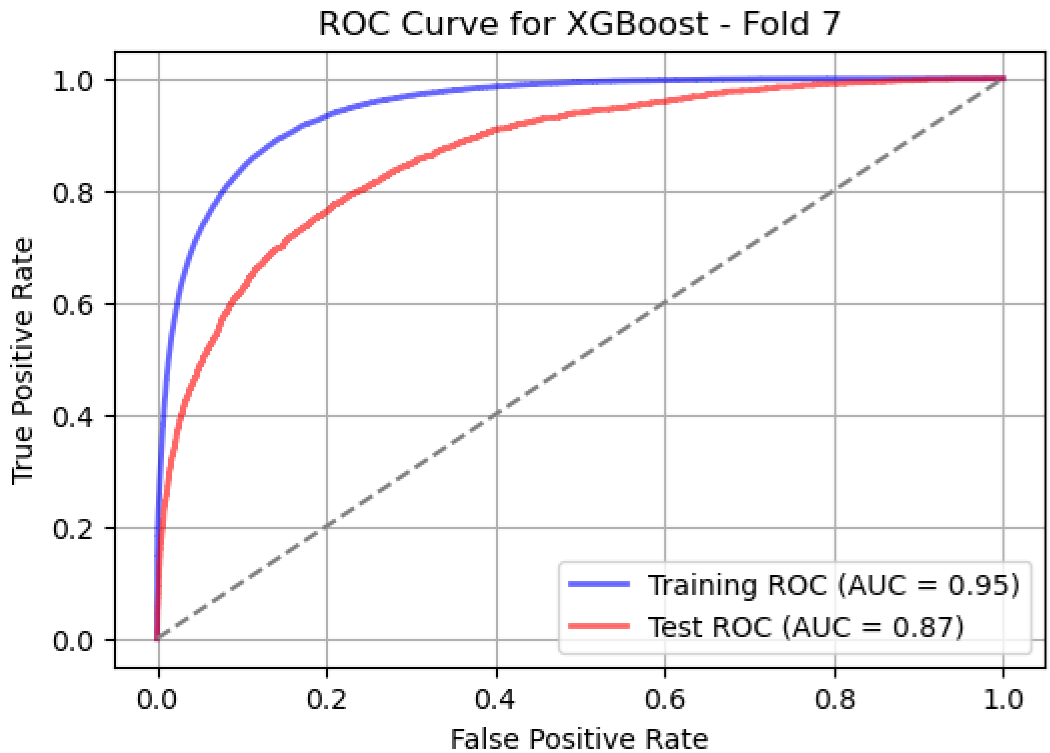

ROC Curve for XGBoost - Fold 6
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Training ROC (AUC = 0.95)
Test ROC (AUC = 0.87)

```
XGBoost Fold 6 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.91
Training F1 Score: 0.53, Test F1 Score: 0.40
Training AUC: 0.95, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.92      0.98      0.95     17990
           1       0.67      0.28      0.40      2010

    accuracy                           0.91     20000
   macro avg       0.80      0.63      0.67     20000
weighted avg       0.90      0.91      0.90     20000

[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 18088, number of negative: 161912
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.438750 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 183117
[LightGBM] [Info] Number of data points in the train set: 180000, number of used fea
tures: 179
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=0.000000
[LightGBM] [Info] Start training from score 0.000000
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
```

```
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

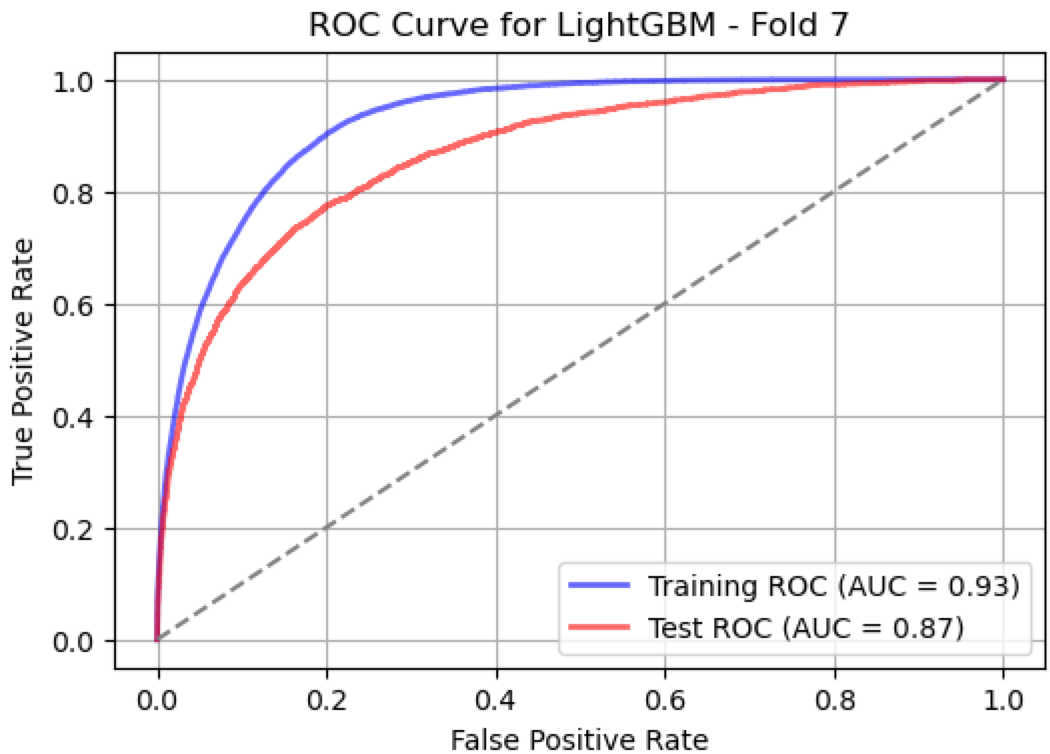



```
LightGBM Fold 6 Metrics:
Training Accuracy: 0.84, Test Accuracy: 0.82
Training F1 Score: 0.52, Test F1 Score: 0.46
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.97      0.83      0.89     17990
           1       0.33      0.74      0.46      2010

    accuracy                           0.82     20000
   macro avg       0.65      0.79      0.68     20000
weighted avg       0.90      0.82      0.85     20000
```

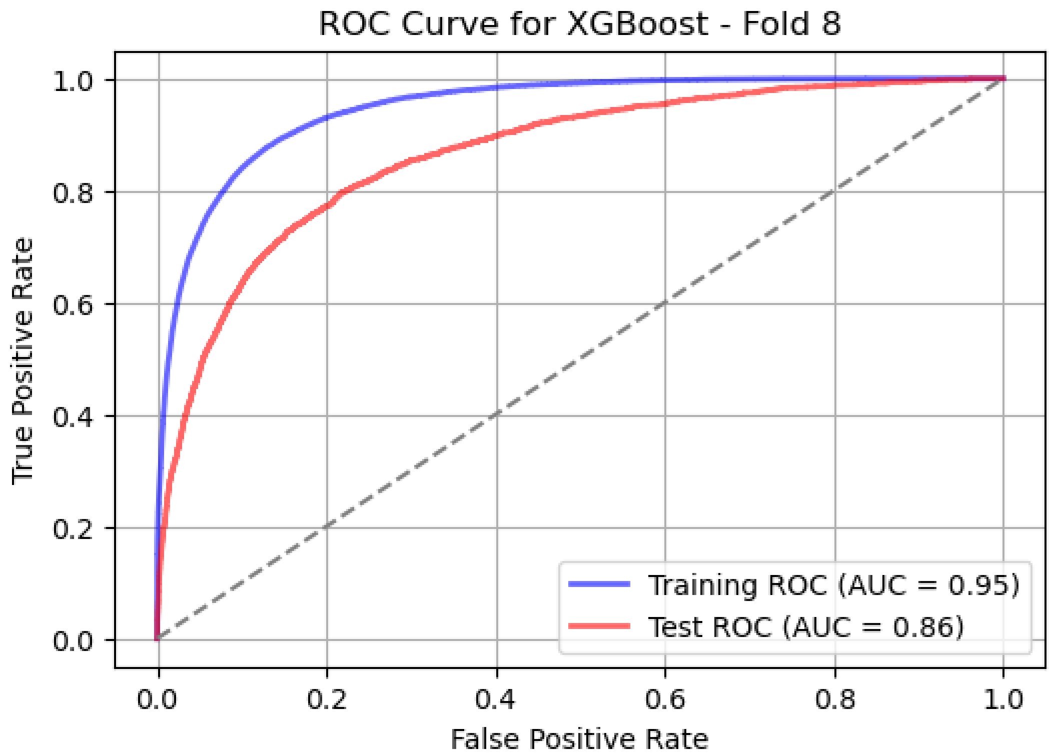

ROC Curve for XGBoost - Fold 7
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Training ROC (AUC = 0.95)
Test ROC (AUC = 0.87)

```
XGBoost Fold 7 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.92
Training F1 Score: 0.53, Test F1 Score: 0.39
Training AUC: 0.95, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     17990
           1       0.71      0.27      0.39      2010

    accuracy                           0.92     20000
   macro avg       0.82      0.63      0.67     20000
weighted avg       0.90      0.92      0.90     20000

[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 18088, number of negative: 161912
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.448420 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 183117
[LightGBM] [Info] Number of data points in the train set: 180000, number of used fea
tures: 179
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=0.000000
[LightGBM] [Info] Start training from score 0.000000
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
```

```
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

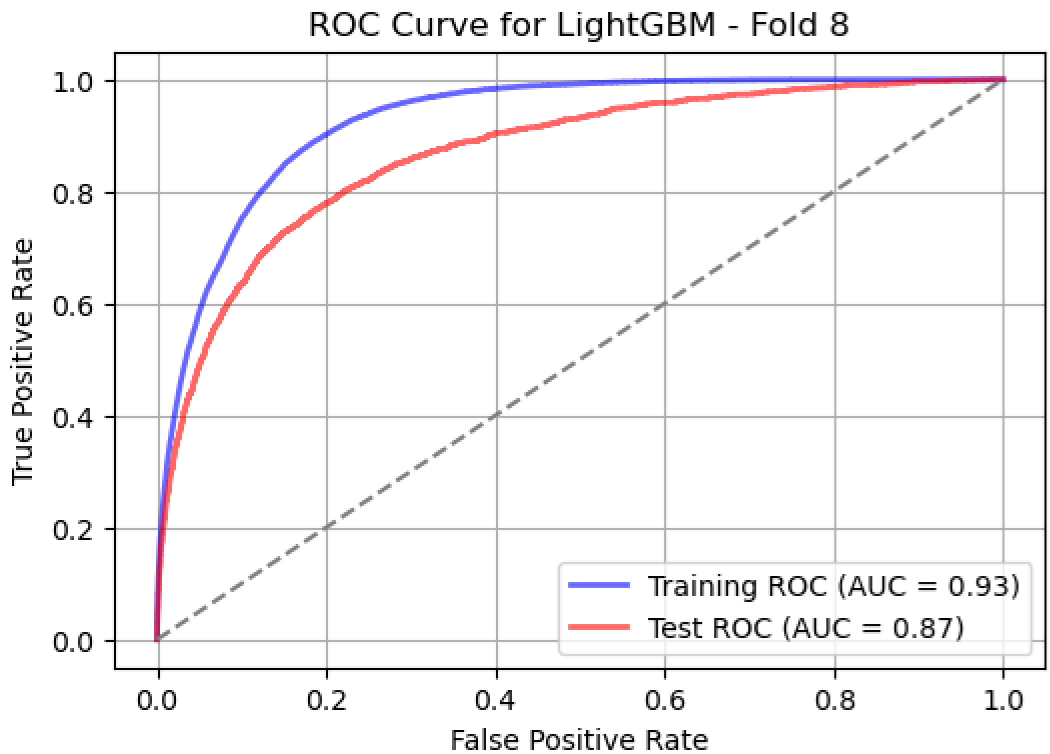



```
LightGBM Fold 7 Metrics:
Training Accuracy: 0.84, Test Accuracy: 0.82
Training F1 Score: 0.52, Test F1 Score: 0.45
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.97      0.83      0.89     17990
           1       0.33      0.74      0.45      2010

    accuracy                           0.82     20000
   macro avg       0.65      0.78      0.67     20000
weighted avg       0.90      0.82      0.85     20000
```

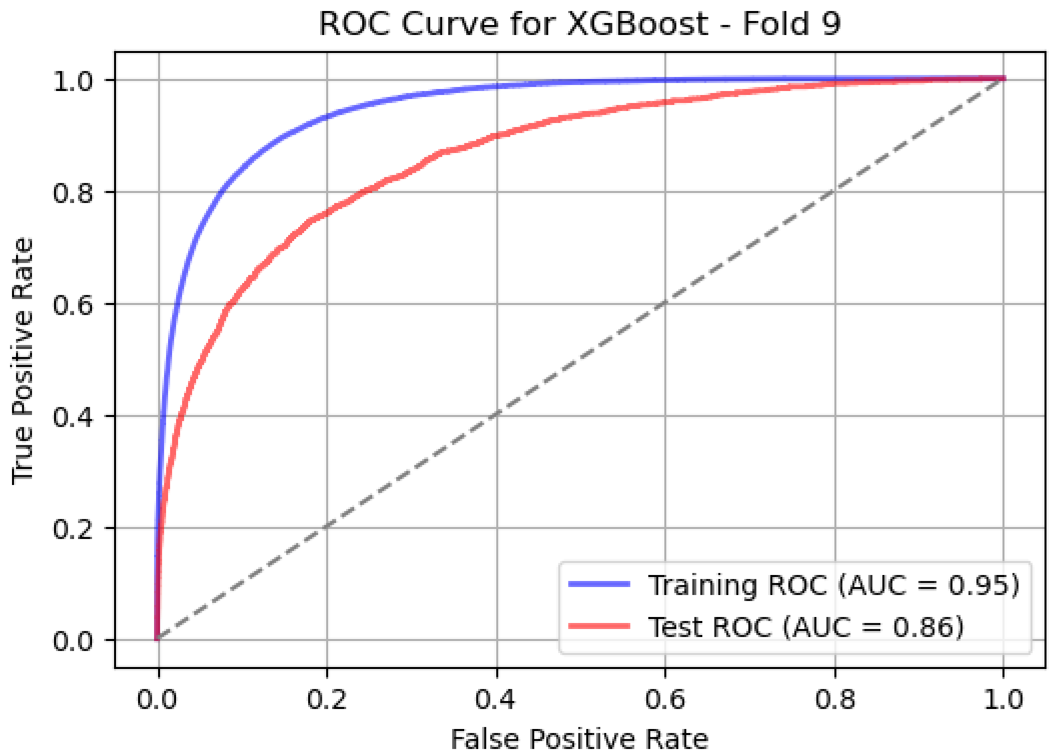

ROC Curve for XGBoost - Fold 8
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Training ROC (AUC = 0.95)
Test ROC (AUC = 0.86)

```
XGBoost Fold 8 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.91
Training F1 Score: 0.53, Test F1 Score: 0.38
Training AUC: 0.95, Test AUC: 0.86
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     17990
           1       0.69      0.27      0.38      2010

    accuracy                           0.91     20000
   macro avg       0.80      0.63      0.67     20000
weighted avg       0.90      0.91      0.90     20000

[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 18088, number of negative: 161912
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.479569 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 183117
[LightGBM] [Info] Number of data points in the train set: 180000, number of used fea
tures: 179
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=0.000000
[LightGBM] [Info] Start training from score 0.000000
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
```

```
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

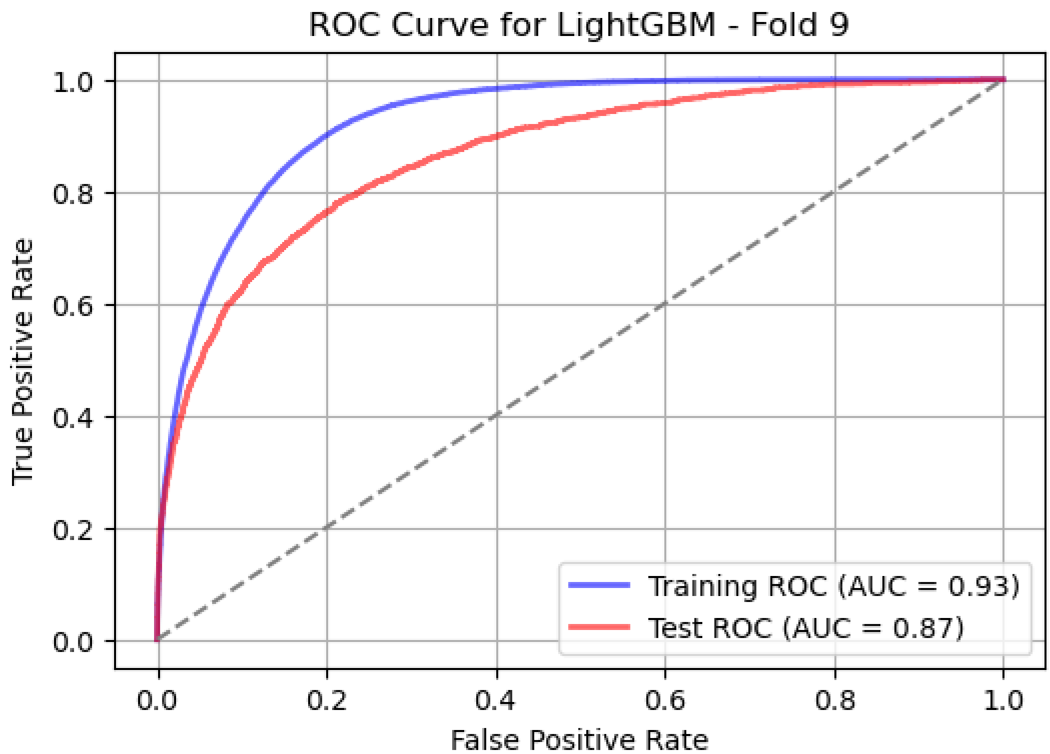



```
LightGBM Fold 8 Metrics:
Training Accuracy: 0.84, Test Accuracy: 0.83
Training F1 Score: 0.52, Test F1 Score: 0.46
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.97      0.84      0.90     17990
           1       0.33      0.74      0.46      2010

    accuracy                           0.83     20000
   macro avg       0.65      0.79      0.68     20000
weighted avg       0.90      0.83      0.85     20000
```

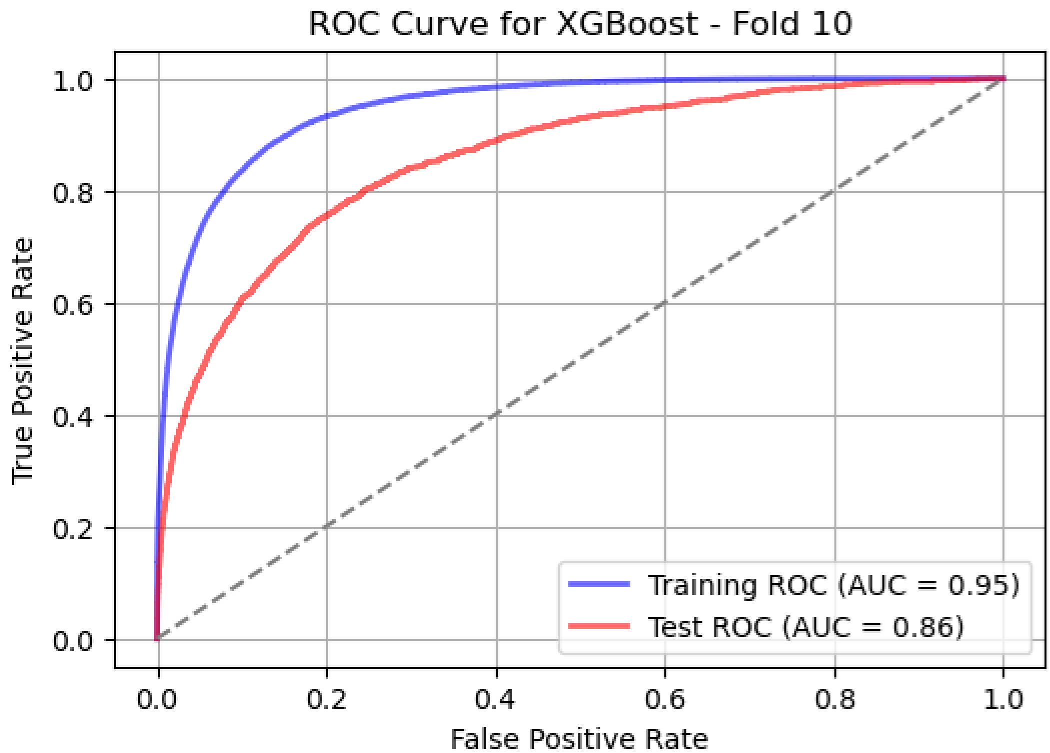

ROC Curve for XGBoost - Fold 9
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Training ROC (AUC = 0.95)
Test ROC (AUC = 0.86)

```
XGBoost Fold 9 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.92
Training F1 Score: 0.53, Test F1 Score: 0.41
Training AUC: 0.95, Test AUC: 0.86
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.92      0.99      0.96     17990
           1       0.72      0.28      0.41      2010

    accuracy                           0.92     20000
   macro avg       0.82      0.64      0.68     20000
weighted avg       0.90      0.92      0.90     20000

[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 18088, number of negative: 161912
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.502708 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 183117
[LightGBM] [Info] Number of data points in the train set: 180000, number of used fea
tures: 179
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=0.000000
[LightGBM] [Info] Start training from score 0.000000
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
```

```
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```

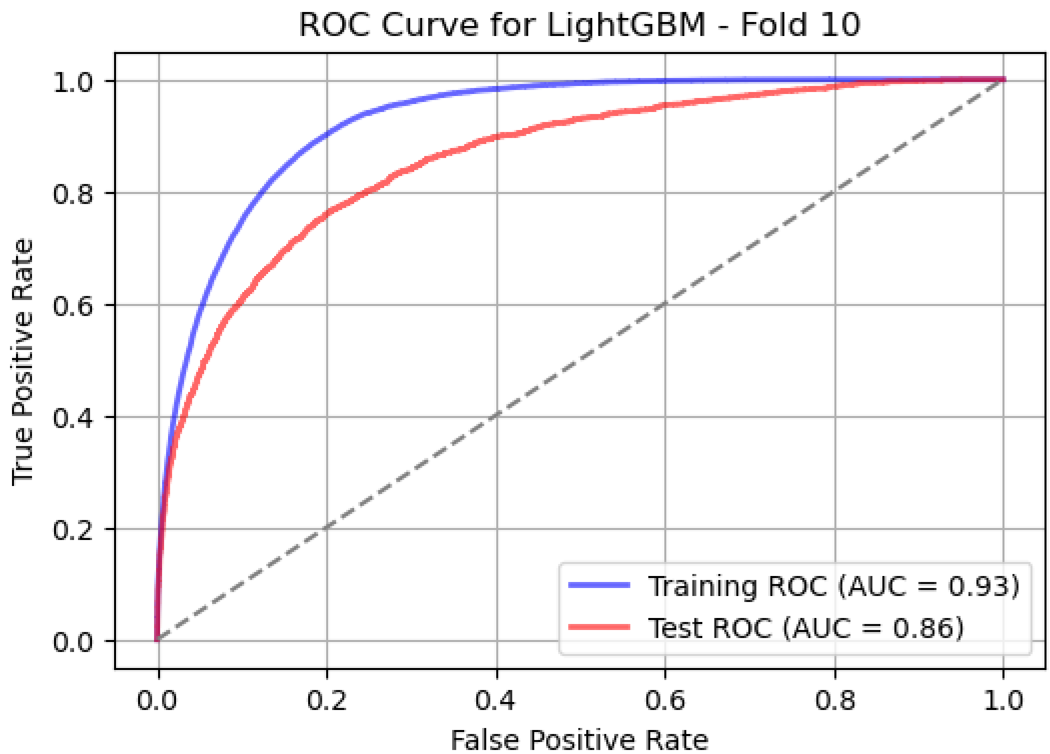



```
LightGBM Fold 9 Metrics:
Training Accuracy: 0.84, Test Accuracy: 0.82
Training F1 Score: 0.52, Test F1 Score: 0.45
Training AUC: 0.93, Test AUC: 0.87
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.96      0.84      0.90     17990
           1       0.33      0.72      0.45      2010

    accuracy                           0.82     20000
   macro avg       0.65      0.78      0.67     20000
weighted avg       0.90      0.82      0.85     20000
```

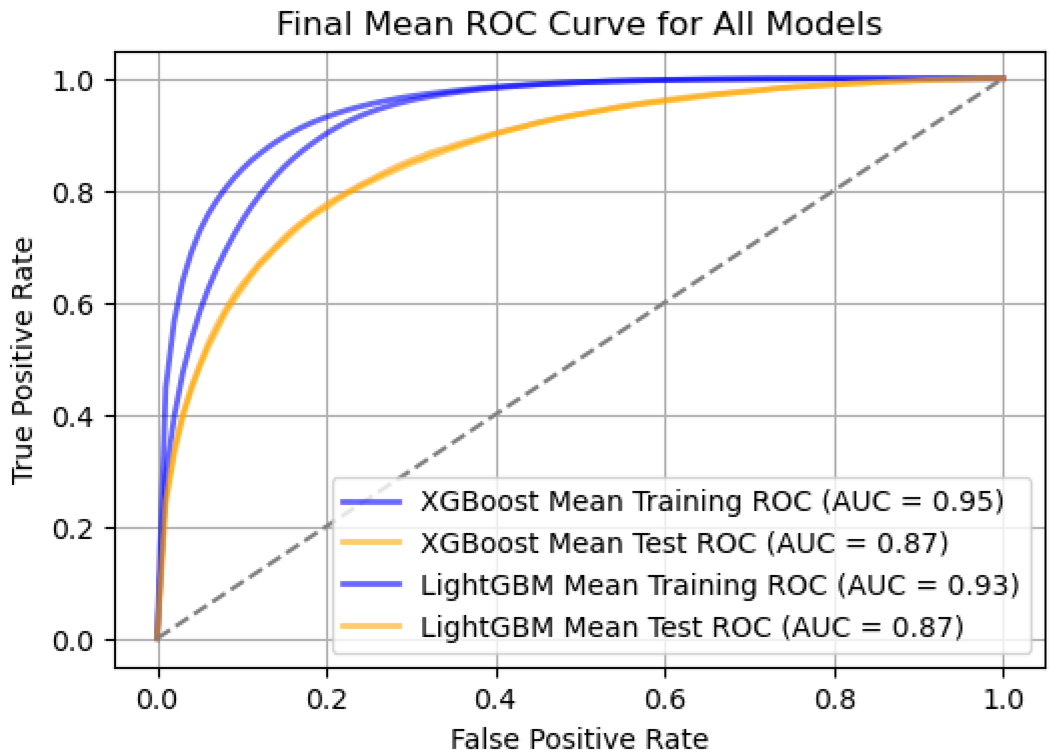

ROC Curve for XGBoost - Fold 10
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Training ROC (AUC = 0.95)
Test ROC (AUC = 0.86)

```
XGBoost Fold 10 Metrics:
Training Accuracy: 0.93, Test Accuracy: 0.92
Training F1 Score: 0.54, Test F1 Score: 0.39
Training AUC: 0.95, Test AUC: 0.86
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     17990
           1       0.71      0.27      0.39      2010

    accuracy                           0.92     20000
   macro avg       0.82      0.63      0.67     20000
weighted avg       0.90      0.92      0.90     20000

[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Info] Number of positive: 18088, number of negative: 161912
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing wa
s 0.401765 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 183117
[LightGBM] [Info] Number of data points in the train set: 180000, number of used fea
tures: 179
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.500000 -> initscore=0.000000
[LightGBM] [Info] Start training from score 0.000000
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignor
ed. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Cu
```

```
rrent value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Curren
t value: bagging_freq=1
```




```
LightGBM Fold 10 Metrics:
Training Accuracy: 0.84, Test Accuracy: 0.82
Training F1 Score: 0.52, Test F1 Score: 0.44
Training AUC: 0.93, Test AUC: 0.86
Classification Report for Test Set:
              precision    recall  f1-score   support

           0       0.96      0.83      0.89     17990
           1       0.32      0.72      0.44      2010

    accuracy                           0.82     20000
   macro avg       0.64      0.77      0.67     20000
weighted avg       0.90      0.82      0.85     20000
```




```
XGBoost Final Cross-Validation Metrics:
Final Accuracy: 0.91
Final F1 Score: 0.39
Final ROC AUC: 0.87

LightGBM Final Cross-Validation Metrics:
Final Accuracy: 0.82
Final F1 Score: 0.46
Final ROC AUC: 0.87
```

Saving Models and Preprocessing Objects

To use the trained models and preprocessing components later, we save them to disk:

In [33]:
```
# Save the models, PCA, and scaler to disk
joblib.dump(models['XGBoost'], r'E:\my_work\dpi\New folder (2)\xgboost_model.pkl')
joblib.dump(models['LightGBM'], r'E:\my_work\dpi\New folder (2)\lightgbm_model.pkl'
joblib.dump(pca, r'E:\my_work\dpi\New folder (2)\pca.pkl')
joblib.dump(scaler, r'E:\my_work\dpi\New folder (2)\scaler.pkl')
```

Out[33]:
```
['E:\\my_work\\dpi\\New folder (2)\\scaler.pkl']
```

Making Test Predictions and Saving Results

The following code cleans the test data, applies scaling and PCA, makes predictions using the trained models, and saves the results:

In [35]:
```
# Function to make predictions on the test set and save the output
def make_test_predictions(models, scaler, pca):
    # Clean the test data (note that test data doesn't have 'target' column)
    test_cleaned = test.drop(columns=['ID_code']).copy()

    # Handle any missing values in test data similarly to training data
    test_cleaned.fillna(test_cleaned.mean(), inplace=True)

    # Apply scaling to the test data
    test_scaled = scaler.transform(test_cleaned)

    # Apply PCA to the scaled test data
    test_pca = pca.transform(test_scaled)

    # Predict probabilities using trained models and save results
    id_codes = test['ID_code']
    results_df = pd.DataFrame({'ID_code': id_codes})

    # Assuming you want to make predictions using all models
    for name, model in models.items():
        predictions = model.predict_proba(test_pca)[:, 1]
        results_df[f'{name}_predicted_probabilities'] = predictions

    # Save the results to a CSV file
    results_df.to_csv(r'E:\my_work\dpi\predictions_with_id.csv', index=False)

    # Display the first few rows of the results for verification
    print(results_df.head())
```

In [36]:
```
make_test_predictions(models, scaler, pca)
```

```
C:\Users\Ahmed\anaconda3\Lib\site-packages\sklearn\base.py:457: UserWarning: X has feature names, but StandardScaler was fitted without feature names
  warnings.warn(
[LightGBM] [Warning] feature_fraction is set=1.0, colsample_bytree=1.0 will be ignored. Current value: feature_fraction=1.0
[LightGBM] [Warning] bagging_fraction is set=0.85, subsample=1.0 will be ignored. Current value: bagging_fraction=0.85
[LightGBM] [Warning] bagging_freq is set=1, subsample_freq=0 will be ignored. Current value: bagging_freq=1
  ID_code  XGBoost_predicted_probabilities  LightGBM_predicted_probabilities
0  test_0                         0.178332                          0.655680
1  test_1                         0.213811                          0.700468
2  test_2                         0.052674                          0.361664
3  test_3                         0.156415                          0.676556
4  test_4                         0.053082                          0.250669
```