LightGBM for GST Data

1. Model Code and Documentation

```
import lightgbm as lgb
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import (accuracy_score, precision_score, recall_score, f1_score,
roc_curve, auc, confusion_matrix, log_loss, balanced_accuracy_score)
import matplotlib.pyplot as plt
import seaborn as sns
Xtrain_data = pd.read_csv('X_Train_Data_Input.csv')
Ytrain_data = pd.read_csv('Y_Train_Data_Target.csv')
Xtest_data = pd.read_csv('X_Test_Data_Input.csv')
Ytest_data = pd.read_csv('Y_Test_Data_Target.csv')
X_train = Xtrain_data.iloc[:, 1::].values
y_train = Ytrain_data.iloc[:,1].values
X_test = Xtest_data.iloc[:, 1::].values
y_test = Ytest_data.iloc[:,1].values
lgb_model = lgb.LGBMClassifier(objective='binary',
boosting_type='gbdt',n_estimators=100,learning_rate=0.1,max_depth=-1,subsample=0.8,
colsample_bytree=0.8, missing=np.nan)
lgb_model.fit(X_train, y_train)
y_pred = lgb_model.predict(X_test)
y_pred_proba = lgb_model.predict_proba(X_test)[:, 1]
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
balanced_acc = balanced_accuracy_score(y_test, y_pred)
logloss_value = log_loss(y_test, y_pred_proba)
fpr, tpr, _ = roc_curve(y_test, y_pred_proba)
roc_auc = auc(fpr, tpr)
cm = confusion_matrix(y_test, y_pred)
print(f"LightGBM Test Accuracy: {accuracy:.4f}")
print(f"LightGBM Test Precision: {precision:.4f}")
print(f"LightGBM Test Recall: {recall:.4f}")
print(f"LightGBM Test F1 Score: {f1:.4f}")
print(f"LightGBM Test Balanced Accuracy: {balanced_acc:.4f}")
print(f"LightGBM Test Log Loss: {logloss_value:.4f}")
print(f"LightGBM Test AUC-ROC: {roc_auc:.4f}")
plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='blue', lw=2, label=f'ROC curve (AUC = {roc_auc:.4f})')
plt.plot([0, 1], [0, 1], color='red', linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.show()
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", cbar=False)
plt.title("Confusion Matrix")
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()
```

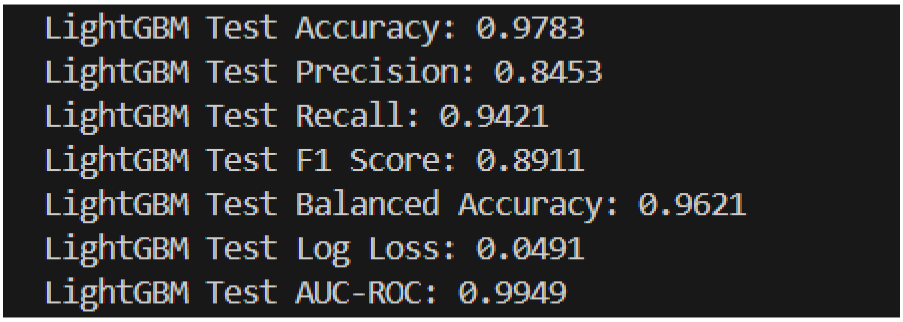
1.1 Steps

- Preprocessing – as a part of feature engineering process the attribute ID (meta data) was exempted from the analysis. Furthermore, data wrangling steps were carried out using pandas library for training and test data.
- Model Execution using lightgbm python library
- Model Evaluation – For evaluation of the model, metrics like precision, recall, accuracy, F1 score, confusion matrix, log loss and AUC-ROC curve has been used.

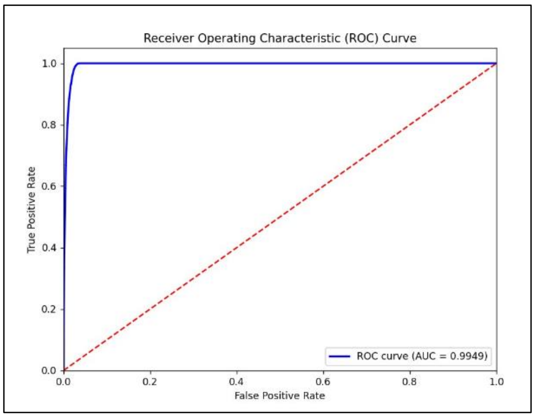
1.2 Why LightGBM?

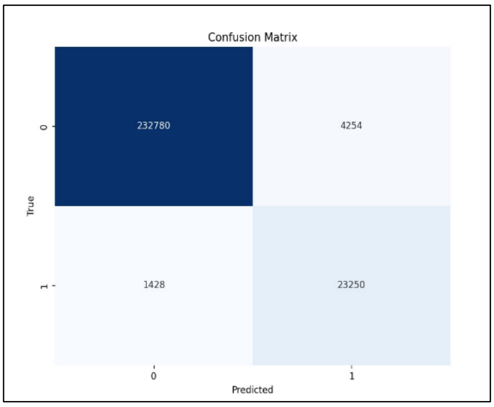
The model that has been used for the classification of the given GST data is LightGBM. LightGBM (Light Gradient Boosting Machine) is an effective and easily accessible gradient boosting framework [1]. It was developed to handle high voluminous data for addressing classification and regression problems. LightGBM combines the prediction of multiple weak decision tree learners in order to improve the accuracy of the model. In contrast to many other tree-based algorithms, which develop trees level-wise, LightGBM grows leaf-wise, resulting in more accurate models and fewer repetitions. LightGBM can handle missing values without explicit imputation. When creating decision trees, it interprets NaN values as a distinct category, allowing it to make decisions depending on their presence. This feature avoids the need for preprocessing techniques such as replacing missing values with mean, median, or mode, which might distort data or create bias. LightGBM ensures the dataset's integrity by allowing the model to learn how to handle NaNs. LightGBM's strong performance and scalability make it popular in a variety of sectors, including finance [2] (for credit scoring), marketing (customer segmentation) , and healthcare (predictive modeling) [] .

2. Model Performance Report



```
LightGBM Test Accuracy: 0.9783
LightGBM Test Precision: 0.8453
LightGBM Test Recall: 0.9421
LightGBM Test F1 Score: 0.8911
LightGBM Test Balanced Accuracy: 0.9621
LightGBM Test Log Loss: 0.0491
LightGBM Test AUC-ROC: 0.9949
```

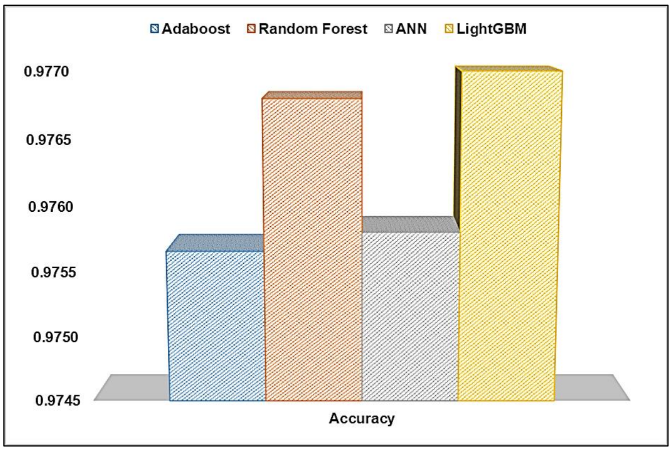




2.2 Performance Analysis

- The model obtained an accuracy of roughly 97.83%, indicating that the model is effective in distinguishing across classes.
- The model's precision of 84.53% means that it accurately predicts positive classes 84.53% of all the time.
- The recall score of 94.21% signifies that the model correctly identifies about 94.21% of the actual positive cases.
- The F1 score is 89.11%. It indicates that the model is not only good at identifying positive cases but also maintains a reasonable level of precision.
- The model possess log loss of 4.97% which is a minimum loss value. Meanwhile, the AUC-ROC is 99.49%, which is the most significant value indicating that the model can be able to correctly classify the positive and negative classes efficiently.

2.3 Benchmarking with state-of-the-art machine learning models



The LightGBM model is benchmarked with state-of-the-art machine learning models such as Adaboost, Random Forest (RF) and Artificial Neural Network (ANN) to establish a clear baseline to analyse the strength and weakness of each model. From the graph, the LightGBM outperforms the state-of-the-art machine learning models.

3. Conclusions

To classify the given GST data, LightGBM model has been used. Since lightGBM can be used for high dimensional data and handles missing values efficiently, the lightGBM model is chosen for the GST data classification problem. LightGBM model have obtained accuracy of 97.83%, precision of 84.53%, recall of 94.21%, F1 score of 89.11%, and AUC-ROC of 0.9949 denoting that the model has performed well in classification of the positive and the negative class. Additionally, the lightGBM model is bench marked with state-of-the-art machine learning algorithms like AdaBoost, Random Forest, and ANN. The benchmark analysis clearly established baseline showing the LightGBM model outperforms on comparing with state-of-the-art models.

References

1. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.
2. Wang, D. N., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. Information Sciences, 602, 259-268.
3. Wang, D., Zhang, Y., & Zhao, Y. (2017, October). LightGBM: an effective miRNA classification method in breast cancer patients. In Proceedings of the 2017 international conference on computational biology and bioinformatics (pp. 7-11).