GST Analytics Hackathon Report: AI and ML Solutions for GST Data

Kaushal Sengupta, Anush Jain, Jay Vivek Yadav, Himani Zambare

Problem Statement

The challenge in this hackathon was to develop a machine learning model capable of accurately predicting the target variable, using the provided GST data. The task is framed as a *binary classification* problem(0 or 1 in this case), where we aim to correctly identify the target based on input features, improving efficiency in GST analysis.

Approach

Our team addressed a binary classification problem, aiming to categorize data into two classes: positive and negative. We constructed a machine learning model implementing **XGBoost (along with threshold and weight adjustment)**, an efficient algorithm, to generate precise predictions. This report outlines our approach to the findings, and the practical applications of this model in real-life scenarios. Our approach involve following steps:

I. **Data Preprocessing**: We tried Imputation of the dataset to fill missing values along with SMOTE method for oversampling the dataset and handle class imbalance and after trying and testing with different approach we found XGBoost was handling missing data own its own and it was not really affecting the performance of our model and also since the exact features of the dataset was not mentioned we decided to keep the integrity and originality of dataset as it is.

II. **Model Selection**: XGBoost was selected due to its superior performance in binary classification problems. We tested other models also such as **Adaboost, Catboost, Fully Connected Neural Network, Random Forest, Naive Bais** but they were also providing similar result. We decided to go with XGBoost as it was giving most promising result with efficiency.

III. **Hyper-parameter Tuning**: We employed the combination of Optuna and RandomizedSearchCV algorithm for optimizing hyperparameters like learning rate, max depth, and the number of estimators.

IV. **Evaluation**: The model was evaluated using metrics such as Accuracy, Precision, Recall, F1 Score, AUC-ROC, and a Confusion Matrix.

Objective

The task given to us was to build a model that predicts a target variable based on input data. To ensure that our model could perform accurately in real-world situations, we evaluated it on multiple metrics, such as accuracy, precision, recall, F1 score, and AUC-ROC. These metrics help us measure how well our model classifies the data.

Methodology

I. Data Preprocessing

As the dataset was already preprocessed and nothing and XGBoost can easily handle missing values we decided not to interfere with it much as unnecessary changes to dataset can deviate us from the result .

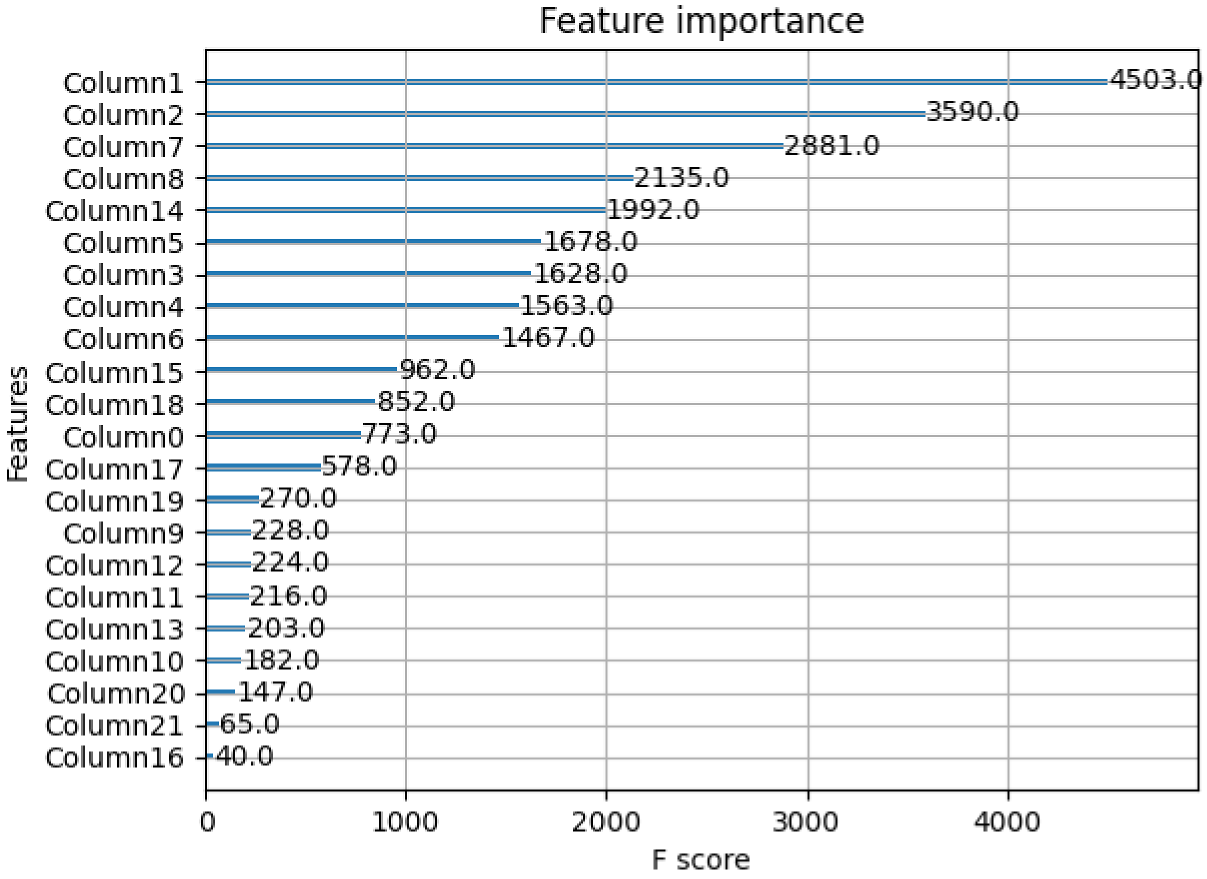

Figure 1: Bar diagram showing the importance of each feature.

II. Model Training and Tuning

We used XGBoost, a robust classification algorithm known for its efficiency and accuracy. We updated the class minority class(1 in this case 74033 of total values of target values) weight to 9.61 times the majority class(0 in this case 711100 of total target values) To fine-tune the model's performance, we employed combination of *Optuna* and *RandomizedSearchCV* algorithm for optimizing hyper-parameters . This process helped us find the best possible combination of model settings.

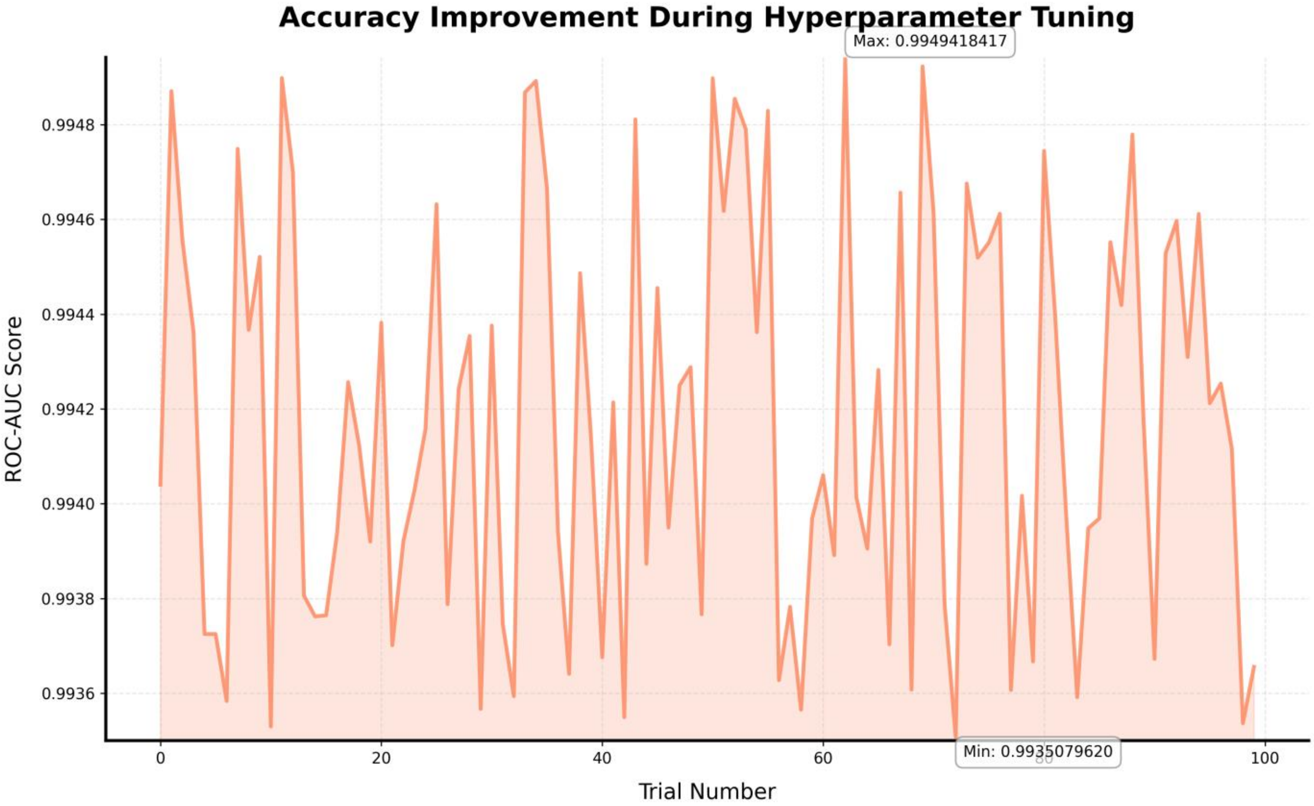

Figure 2: Graph showing accuracy improvement during hyper-parameter tuning.

Below is the best found parameters for our model with ROC-AUC score of **0.9949418417329653** are:

subsample: **0.8**

n_estimators: **500**,

min_child_weight: **7**,

max_depth: **7**,

learning_rate: **0.05**,

lambda: **0**,

gamma: **0.1**,

colsample_bytree: **0.9**,

alpha: **0**

Evaluation Metrics

To evaluate how well our model works, we used the following key metrics:

I. Accuracy

This metric tells us how often the model makes the right prediction. An accuracy score of 80% means that the model is correct 80 out of 100 times.

II. Precision

Precision measures the percentage of correctly identified positives out of all the predictions the model made.

III. Recall (Sensitivity)

Recall tells us how good the model is at identifying positive cases. This metric becomes essen tial when missing a positive case can have serious consequences.

IV. F1 Score

The F1 score is a balanced measure that considers both precision and recall. It provides a single number to summarize performance when both metrics are equally important.

V. AUC-ROC

AUC-ROC is a graphical measure of how well our model distinguishes between the two classes. The closer the value is to 1, the better the model performs.

VI. Confusion Matrix

The confusion matrix provides a detailed breakdown of correct and incorrect predictions. It includes the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Results and Insights

- **Accuracy**: Our model achieved an accuracy of 97.84% on test data, meaning it correctly predicted the target variable in 97.84 out of 100 cases.
- **Precision**: We achieved a precision score of 0.8496, meaning that 84.96% of the positive predictions were correct.
- **Recall**: With a recall score of 0.9394, our model successfully captured 93.94% of the actual positive cases.
- **F1 Score**: The F1 score of 0.8923 indicates a balanced performance between precision and recall.
- **AUC-ROC**: Our model obtained an AUC score of 0.9949, showcasing excellent performance in distinguishing between the two classes.

- **Confusion Matrix**: The confusion matrix showed that the model made only a few incorrect predictions, with the majority of cases being accurately classified.

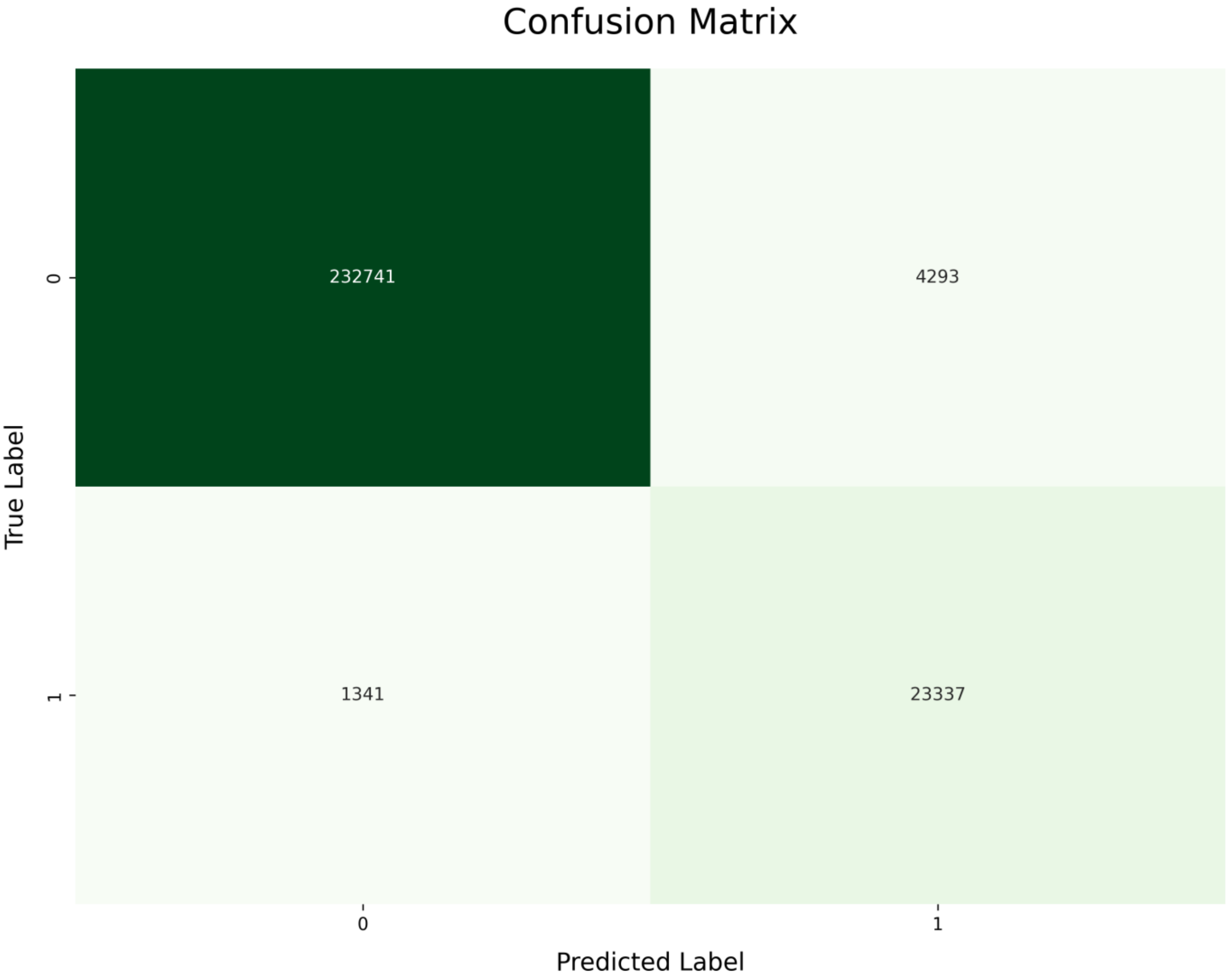

Figure 5: Confusion matrix heat map showing classification results.

Impact and Conclusion

Our solution leverages state-of-the-art machine learning techniques to provide a reliable and robust classification model for GST data. By accurately predicting the target variable, the model can assist in improving the efficiency of the GST analytics framework, aiding government agencies in detecting anomalies or patterns that require further investigation.

The use of hyperparameter optimization has further enhanced the model's performance, making it suitable for real-world application where accuracy and reliability are crucial. Our approach demonstrates the potential of AI and ML in enhancing tax data analysis, benefiting the overall GST system.

Citation Report

- **XGBoost Documentation**: XGBoost Documentation (https://xgboost.readthedocs.io/en/latest/)

- **Scikit-learn Documentation**: Scikit-learn Documentation(https://scikit-learn.org/stable/)

- **Pandas Documentation**: Pandas Documentation(https://pandas.pydata.org/docs/)