# **Report on Fraud Detection in GST Transactions Using Machine Learning Techniques**

GitHub: <https://github.com/Aakashdeep-Srivastava/GST-Hackathon>

Google Colab : <https://colab.research.google.com/drive/1SnItLX3-yvKVkwVRfNin3Oo9GsPEGGFg?usp=sharing>

## **1. Introduction**

### **1.1 Background**

Fraud detection in Goods and Services Tax (GST) is a critical issue in India due to the increasing incidence of fraudulent activities, including fake invoices, input tax credit (ITC) fraud, and underreporting of sales. These activities not only lead to significant revenue losses for the government but also undermine the integrity of the tax system. Identifying fraudulent transactions poses a considerable challenge, particularly in a system characterized by a high volume of transactions, complex data patterns, and a low incidence of fraudulent cases. Traditional methods of fraud detection often fall short due to their inability to efficiently handle large datasets and recognize the nuanced patterns indicative of fraudulent behavior. In recent years, machine learning techniques have emerged as promising solutions, offering enhanced capabilities for detecting anomalies and identifying potential fraud.

### **1.2 Problem Statement**

The dataset under consideration for this study consists of 21 features, including the target variable indicating fraudulent or non-fraudulent transactions. One of the primary challenges encountered is the class imbalance within the dataset; fraudulent cases constitute a very small percentage of the total transactions. This imbalance complicates the prediction of fraudulent activities, as standard classification algorithms may perform poorly on underrepresented classes. To address this issue, we aim to develop a custom mathematical model designed specifically to handle class imbalance and accurately predict fraudulent GST transactions.

### **1.3 Objective**

The objectives of this study are as follows:

* To construct a predictive model capable of accurately detecting fraudulent GST transactions.
* To implement techniques that address the imbalance in the dataset using a custom mathematical approach.
* To evaluate model performance using metrics tailored for imbalanced data, ensuring a comprehensive assessment of the model's effectiveness.

## **2. Approach**

To achieve the stated objectives, the following approach was undertaken:

### **2.1 Data Collection and Preprocessing**

* **Data Acquisition**: The dataset was collected from relevant GST records, comprising various transaction features.
* **Data Cleaning**: Missing values, duplicates, and inconsistent entries were addressed to ensure data quality.
* **Feature Engineering**: New features were created to enhance the model's predictive power, and categorical variables were encoded appropriately.

### **2.2 Addressing Class Imbalance**

* **Resampling Techniques**:
  + **Oversampling**: Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) were applied to generate synthetic instances of the minority class.
  + **Undersampling**: Random undersampling of the majority class was considered to balance the dataset.
* **Cost-sensitive Learning**: Modifications were made to the classification algorithms to assign higher misclassification costs to the minority class, enhancing the model's sensitivity to fraud.

### **2.3 Model Selection and Training**

* Several machine learning algorithms were evaluated, including:
  + Logistic Regression
  + Decision Trees
  + Random Forest
  + XGBoost
  + LightGBM
* Models were trained using stratified k-fold cross-validation to ensure a robust evaluation of performance across various subsets of the data.

### **2.4 Evaluation Metrics**

* Given the class imbalance, traditional metrics such as accuracy were supplemented with:
  + Precision
  + Recall
  + F1-score
  + ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

These metrics provided a more nuanced understanding of model performance, especially in terms of detecting fraudulent cases.

## **3. Results**

### **3.1 Model Performance**

The results obtained from the different models were as follows (insert results here):

**Overall accuracy and precision , recall , F1 score of minority class 1 and ROC-AUC (%)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model/Approach** | **Accuracy (%)** | **Precision(%)** | **Recall (%)** | **F1-Score (%)** | **ROC-AUC (%)** |
| **Base Models** |  |  |  |  |  |
| XGBoost | 0.98 | 0.85 | 0.94 | 0.89 | 0.99 |
| Gaussian Naive Bayes | 0.96 | 0.74 | 0.84 | 0.79 | 0.96 |
| **Under-Sampling** |  |  |  |  |  |
| XGBoost (Under-Sampling) | 0.97 | 0.76 | 1.00 | 0.86 | 0.99 |
| Gaussian Naive Bayes (Under-Sampling) | 0.96 | 0.75 | 0.79 | 0.77 | 0.96 |
| **Over-Sampling** |  |  |  |  |  |
| XGBoost (Over-Sampling) | 0.97 | 0.78 | 0.99 | 0.87 | 0.99 |
| Gaussian Naive Bayes (Over-Sampling) | 0.96 | 0.73 | 0.89 | 0.80 | 0.96 |
| **SMOTE** |  |  |  |  |  |
| XGBoost (SMOTE) | 0.98 | 0.83 | 0.95 | 0.89 | 0.99 |
| Gaussian Naive Bayes (SMOTE) | 0.92 | 0.54 | 0.88 | 0.67 | 0.94 |

### **3.2 Comparison and Analysis**

The performance metrics indicated that **XGBoost consistently outperformed Gaussian Naive Bayes across various methodologies**. In particular, both base models achieved high accuracy, but XGBoost demonstrated superior precision, recall, and F1-score metrics, particularly in the under-sampling and SMOTE approaches. The ability of the XGBoost model to achieve a recall of 1.00 in the under-sampling scenario suggests its effectiveness in identifying all fraudulent transactions in the validation set, which is crucial given the low incidence of fraud in the dataset.

The implemented resampling techniques significantly enhanced the models' ability to detect fraudulent transactions. For instance, under-sampling increased the recall for the XGBoost model to 1.00, indicating that it successfully identified all instances of the minority class. Similarly, SMOTE improved the performance of the XGBoost model, resulting in a precision of 0.83 and maintaining a high recall of 0.95, showcasing the benefits of addressing class imbalance.

### **4. Conclusion**

In conclusion, the study successfully developed a predictive model for fraud detection in GST transactions, addressing the critical issue of class imbalance. The implemented machine learning techniques demonstrated considerable promise in accurately identifying fraudulent activities, which is crucial for safeguarding revenue and maintaining the integrity of the GST system in India. The use of resampling techniques like under-sampling and SMOTE proved effective in improving model performance, particularly in terms of recall and F1-score, thereby enhancing the overall predictive capabilities of the models. Future work can explore further optimization and the integration of more advanced algorithms to enhance detection accuracy even further.

### **4.1 Future Work**

Future research could explore advanced techniques such as deep learning or ensemble methods to further enhance model performance. Additionally, the integration of real-time data and feedback loops could help in adapting the model to evolving fraud patterns, ensuring continuous improvement in detection capabilities.