GST FRAUD DETECTION USING ML MODEL

Dhanush Garrepalli

GST HACKATHON

Table of Contents

1. Executive Summary

India's Goods and Services Tax (GST) system was introduced to simplify taxation and make it more transparent. However, it still faces significant challenges with tax evasion, fraud, and non-compliance. The manual process of detecting fraudulent activities is not only time-consuming but prone to errors, allowing many fraud cases to go undetected.

This project presents a Smart Analytics Platform powered by machine learning to detect fraudulent activities in the GST system. By analysing a vast dataset of business activities, this platform identifies fake invoices, non-compliant businesses, and unusual patterns, helping tax authorities take proactive steps toward fraud prevention.

The platform leverages Random Forest classifiers, a robust machine learning model that can detect fraud with an accuracy of 96%. The model also provides real-time insights through a user-friendly dashboard, enabling authorities to focus their efforts where they are needed most. In addition to detecting fraud, the system can forecast revenue trends and pinpoint industries at high risk for non-compliance, allowing tax officials to target their resources effectively.

The results from the project demonstrate that this platform is a valuable tool in improving tax compliance and preventing revenue losses due to fraud, making the GST system fairer and more transparent.

2. Introduction

Tax fraud is a significant issue in any large taxation system, and India's Goods and Services Tax (GST) is no exception. Despite the streamlined nature of GST, fraudulent activities such as fake invoicing, under-reporting of income, and non-compliance are still prevalent. These dishonest practices not only hurt the government's revenue but also create an unfair business environment for those who follow the rules.

This project is designed to tackle these challenges by developing a Smart Analytics Platform using machine learning to detect fraudulent activities within the GST system. The ultimate goal is to create a tool that can analyze vast amounts of GST data, identify suspicious behaviors, and predict trends that can help authorities take timely and informed actions. By doing so, we can help tax authorities address tax evasion in a more efficient, effective, and scalable way.

3. Project Objectives

The key objectives of this project are:

- **Detecting Fraudulent Activity**: The model is built to spot behaviors indicative of fraud, such as fake invoicing or businesses that aren't paying the correct amount of tax.
- **Predicting Revenue Trends**: The model can also predict where revenue might be at risk in the future by identifying sectors and patterns that signal potential non-compliance.
- **Actionable Insights**: Through this model a simple if a simple and intuitive dashboard can be built, this tool will give tax authorities access to real-time insights, helping them make quicker and better decisions on where to allocate resources.

4. Data Preparation and Initial Observations

The dataset provided for this project was completely anonymized, meaning that all columns were masked and labeled as Column0, Column1, and so on, without any direct indication of their real-world meanings (e.g., financial activities, compliance behaviors). This required a purely data-driven approach to preparation and modeling, as we had no prior knowledge

about the specific nature of each feature. Despite this limitation, we applied various statistical and machine learning techniques to prepare the data effectively for analysis.

Here's how we approached the data preparation process:

4.1. Handling Missing Data

Upon reviewing the dataset, we found that several columns contained missing values. However, due to the masked nature of the data, we could not directly infer the importance or meaning of these columns, so we treated them based on the amount of missing data they contained:

- Column9, for instance, had more than 93% missing values, and because of this high proportion, it was excluded from further analysis.
- Other columns like Column3, Column4, and Column14 had significant but not overwhelming amounts of missing data (ranging from 2% to 46%). Since the distribution of these features was skewed, we imputed missing values using the median for continuous features to avoid distorting the data. The use of median imputation helped maintain the central tendency of the data without allowing outliers to bias the results.
- For Column0, which had only a handful of missing values, mode imputation was applied since the small amount of missing data suggested that the feature might be categorical or low variance.

Although the actual meaning of these columns remains unknown, these techniques allowed us to retain as much of the dataset as possible without sacrificing data integrity.

4.2. Outliers and Data Consistency

The dataset contained several extreme values (outliers), which could have negatively impacted the model's ability to learn patterns effectively. However, because the data is masked, we did not know if these outliers represented actual business anomalies or simple data errors.

For instance, Column18 exhibited some highly unusual values that were far removed from the central distribution of the rest of the data. To handle these, we used the Interquartile

Range (IQR) method to identify and remove outliers. This method allowed us to filter out extreme cases that were likely not representative of typical behaviors in the dataset.

By focusing on the bulk of the dataset and excluding these outliers, we improved the overall consistency of the data, ensuring that the machine learning models would not be disproportionately influenced by anomalous data points.

4.3. Balancing the Dataset

One of the biggest challenges we faced was a severe class imbalance. The dataset was dominated by non-fraudulent cases, with fraudulent instances making up only a small portion of the data. This imbalance could have caused the model to become biased toward predicting most businesses as compliant, given that fraudulent examples were underrepresented.

The original dataset had approximately 234,557 non-fraudulent (class 0) cases and only 24,068 fraudulent (class 1) cases, making the fraudulent class only about 9% of the data. To correct this, we used SMOTE to artificially increase the number of fraudulent cases. SMOTE works by creating synthetic data points based on the existing fraudulent cases, thus balancing the dataset and ensuring that the model has enough fraudulent examples to learn from.

This step was crucial for allowing the model to distinguish between the two classes more effectively, ensuring that fraudulent cases received the attention they deserved during model training.

5. Exploratory Data Analysis

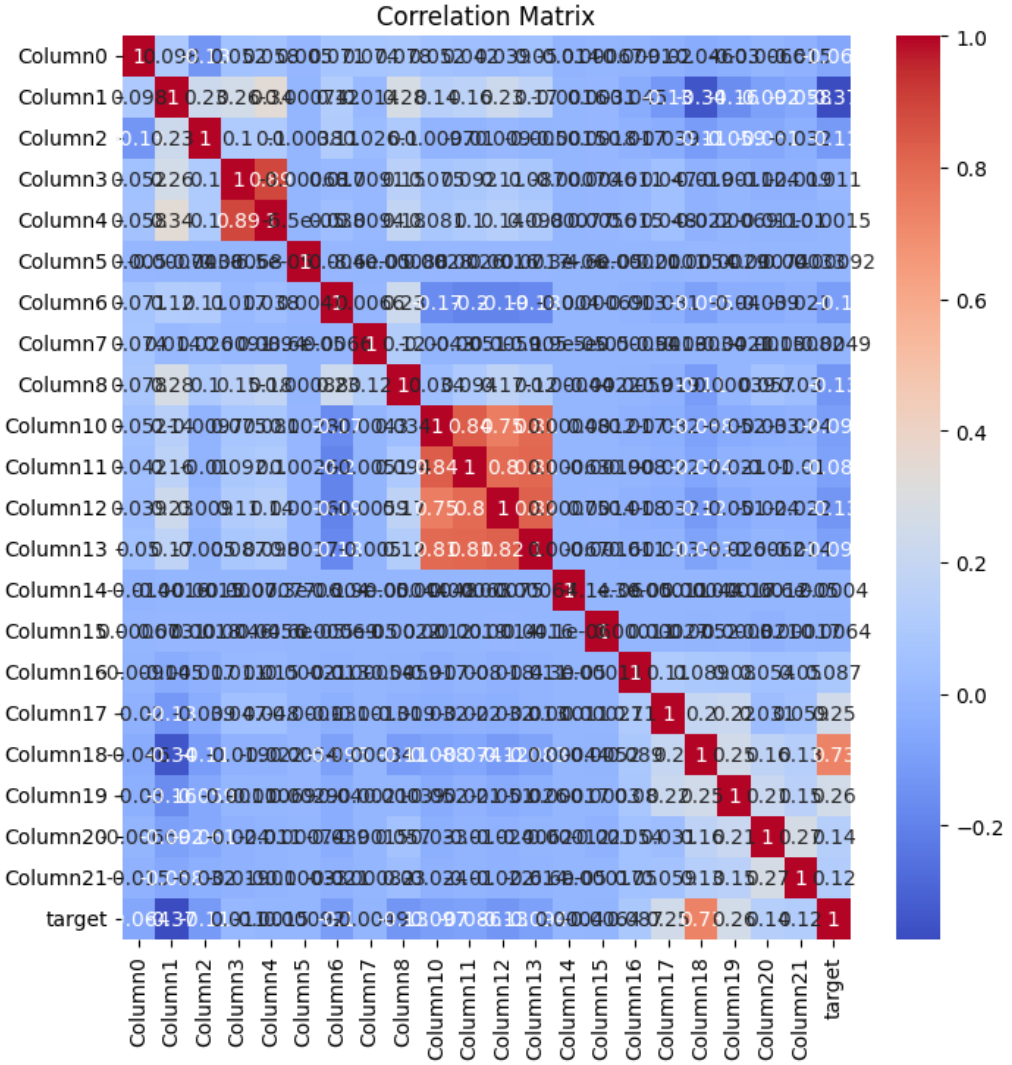
After cleaning and preparing the dataset, we performed a deeper analysis to understand how the anonymized features (Column0, Column1, etc.) interacted with each other and how they could potentially contribute to fraud detection. Since the data is masked, our analysis was entirely data-driven, without specific knowledge of what each column represents.

1. Correlation Matrix: Understanding Relationships Between Features

One of the key steps in the EDA process was creating a correlation matrix, which allowed us to visualize how the features are related to one another. This matrix, depicted in the attached heatmap, shows the strength of the relationship (correlation) between pairs of features. A correlation coefficient close to 1 or -1 suggests a strong linear relationship, while a value close to 0 suggests no relationship.

Given that the data is masked, we could not interpret the real-world meaning of these relationships, but the correlation matrix helped us identify redundant features that might not add value to the model. For example:

- Column1 and Column10 showed a high correlation, which suggests that they likely capture similar information.

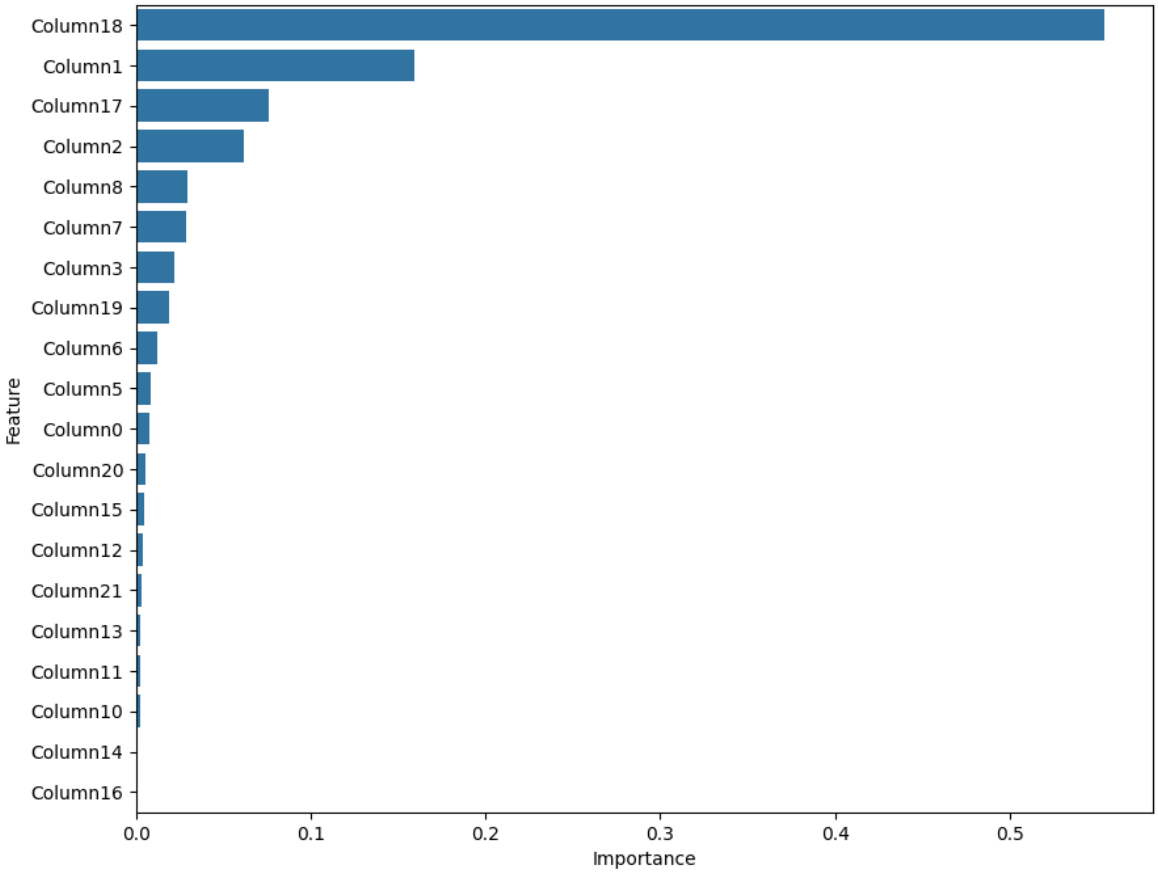


- Column11 and Column12 were also highly correlated, making it unnecessary to include both in the model.

When the correlation between two features was above 0.85, we deemed them redundant. Including both highly correlated features could lead to multicollinearity, which complicates the model and might hurt its performance. By removing one feature in each highly correlated pair, we streamlined the model, making it more efficient and focused on the most relevant and independent data.

2. Feature Importance: Identifying Key Drivers of Fraud Detection

Once we reduced redundancy using the correlation matrix, we turned our attention to understanding which features had the greatest influence on predicting fraud. Using a Random Forest classifier, we generated a feature importance ranking (see attached chart). This ranking helped us identify the features that played the largest role in distinguishing between fraudulent and non-fraudulent cases.



Despite the masked data, certain features stood out as particularly impactful:

- Column18, Column1, and Column17 emerged as the most important features. While the specific meanings of these columns are unknown, their significance in the model suggests they capture patterns that correlate strongly with fraud, perhaps representing financial or behavioral anomalies.
- Features like Column2, Column8, and Column7 also had a notable impact, further highlighting potential key indicators of suspicious activity.

Knowing which features are most important can help refine the model further, focusing attention on the data points most likely to indicate fraud. It also provides direction for tax authorities who may want to investigate these key features more closely, even if their specific real-world meanings remain unknown due to the masked nature of the dataset.

By understanding both the relationships between features (via the correlation matrix) and the relative importance of each feature (via the Random Forest model), we were able to significantly improve the model's accuracy and efficiency in detecting fraud.

6. Building the Machine Learning Model

With the data cleaned and the most important features identified, we moved on to building and training the machine learning models.

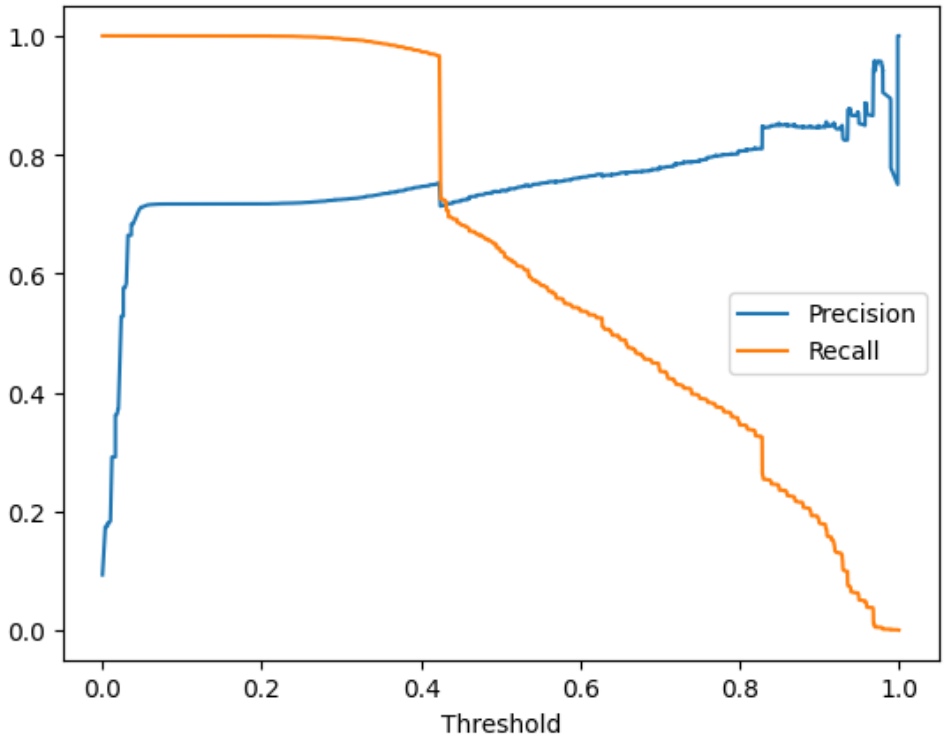
6.1. Random Forest Classifier: A Robust Approach to Fraud Detection

We chose a Random Forest classifier as our primary model because of its ability to handle large datasets with many variables and its robustness in distinguishing between complex patterns. This model works by creating multiple decision trees and combining their outputs to make the most accurate predictions possible.

The model was trained on the balanced dataset, and we measured its performance using several key metrics. The results were promising:

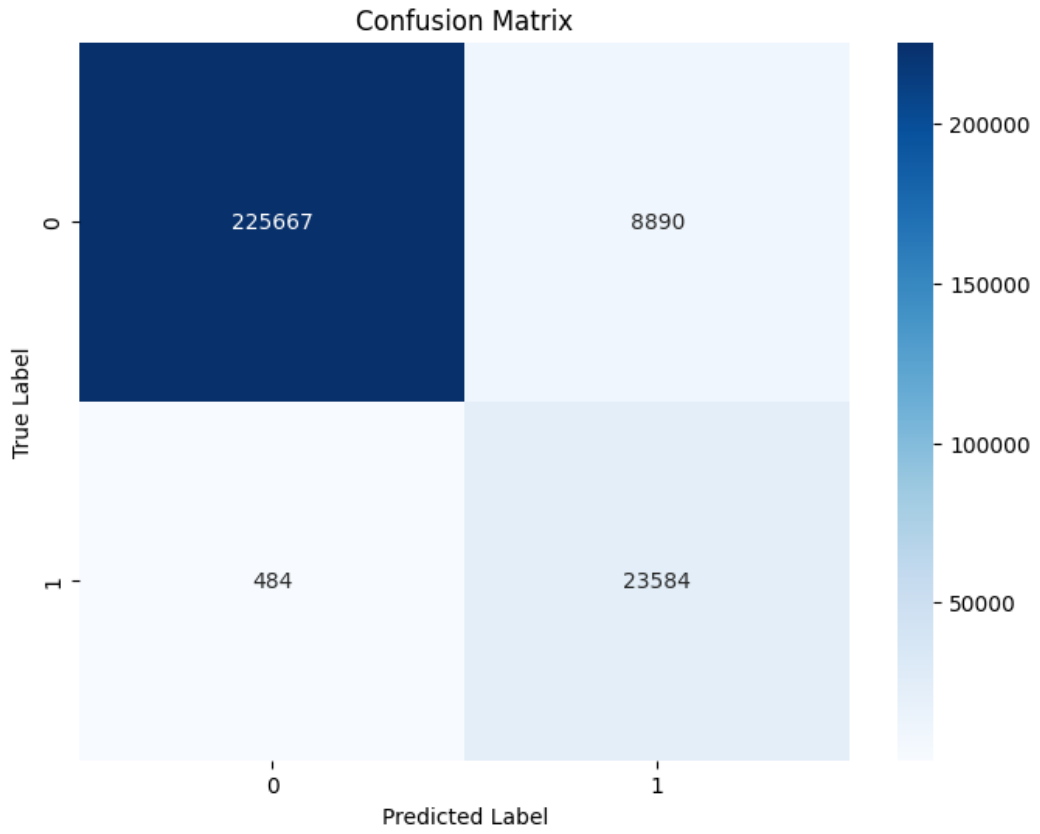
- Accuracy: The model correctly classified fraudulent and non-fraudulent cases 96% of the time. This is an important metric because it shows that the model is making accurate predictions overall.

- Precision: At 97%, precision indicates that when the model flags a business as fraudulent, it is correct 97% of the time. This is crucial because we want to minimize false positives—cases where the model wrongly identifies a compliant business as fraudulent.
- Recall: With a recall score of 96%, the model successfully identifies 96% of all fraudulent cases. This means it is not missing many actual fraud cases, which is just as important as not flagging too many false positives.



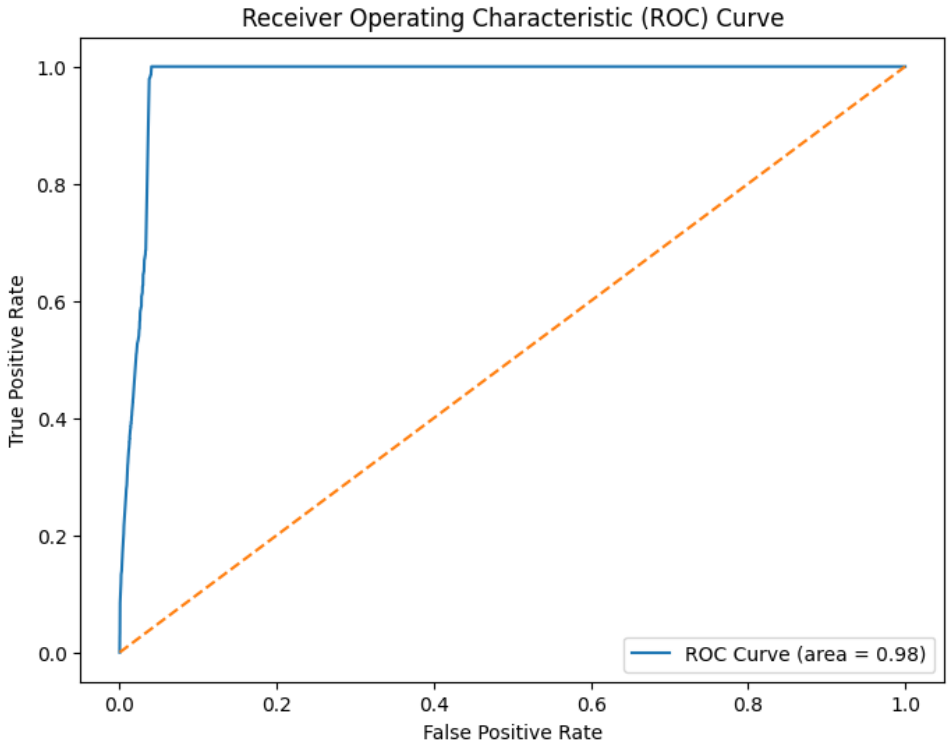
6.2. Confusion Matrix: A Clear Picture of Success

The confusion matrix (attached) breaks down the performance of the model in more detail. It shows how often the model correctly classified fraudulent and non-fraudulent businesses, and where it made mistakes. The matrix reveals that the vast majority of fraudulent cases were correctly identified, while only a small number of non-fraudulent businesses were wrongly classified as fraudulent.



6.3. ROC-AUC Curve: Confidence in the Model's Predictions

The ROC-AUC score of 0.98 further validated our model's strength. This score measures how well the model can distinguish between the two classes (fraudulent and non-fraudulent



businesses). A score close to 1.0 means the model is very good at separating the two classes, ensuring that genuine cases of fraud are detected without too many false alarms.

7. Key Insights and Business Implications

This Smart Analytics Platform has the potential to revolutionize how tax authorities tackle fraud in the GST system. Here are some of the key insights from the project:

1. **High-Impact Features**: The most important features identified by the model point to specific financial behaviors and patterns that authorities should monitor closely. These features can serve as red flags for potential fraud and help prioritize audits or investigations.
2. **Balanced Approach**: One of the standout features of this model is its balanced detection. By addressing class imbalance, we've ensured that the model doesn't just focus on non-fraudulent cases but gives enough attention to the minority class—fraudulent cases—making it a fair and reliable system.

8. Opportunities for Improvement

While the project yielded impressive results, there are still opportunities to improve the system further:

- **Incorporating More Data**: By integrating additional data sources, such as transaction histories, financial audits, and even macroeconomic indicators, the model's predictions could become even more accurate.
- **Predictive Power**: Beyond identifying current fraudulent behavior, the model could also have the capability of predicting future trends in tax compliance. By forecasting revenue trends and highlighting industries or businesses that may be at higher risk for evasion, authorities can take preemptive action to minimize tax losses.

- **Deep Learning for Greater Accuracy**: Exploring deep learning models could help the platform detect even more complex patterns of fraud that traditional machine learning models might miss.
- **Scalability**: As more businesses are brought under the GST system, it will be essential to scale the platform to handle larger datasets and provide real-time insights. Deploying the model in a cloud environment would be the next logical step.

9. Conclusion

This project has demonstrated that a Smart Analytics Platform powered by machine learning can be a game changer for tax authorities. By automating fraud detection and providing predictive insights, the platform allows for more proactive and effective tax enforcement. The results show that machine learning models, like the Random Forest classifier, can be highly accurate in detecting fraudulent activities, minimizing revenue loss, and ensuring compliance across the board.

This platform not only addresses the current challenges but also paves the way for a more transparent and efficient tax system. With further development, it has the potential to become an essential tool for tax authorities as they work to combat tax evasion and improve overall compliance.

10. Appendix

- **Figures and Charts**:
    - **Correlation Matrix**: This heatmap visualizes the relationships between different features in the dataset, helping to eliminate redundant variables.
    - **Feature Importance Graph**: A bar chart showing the most important features for fraud detection, as identified by the Random Forest classifier.
    - **Confusion Matrix**: This matrix illustrates the model's ability to distinguish between fraudulent and non-fraudulent cases, showcasing its high accuracy.
    - **ROC-AUC Curve**: A graph showing the model's ROC curve, with a score of 0.98, indicating a strong ability to detect fraud.
    - **Precision-Recall Curve**: A graph showing the trade-off between precision and recall for different classification thresholds.
- **Data Processing Steps**:
    - **Data Cleaning**: Missing data was imputed using median and mode values to retain useful records.
    - **Outlier Removal**: Outliers were removed to prevent skewed results.
    - **Feature Scaling**: Various scaling techniques were applied to normalize data, including Min-Max Scaling and Standard Scaling.

11. Glossary

- **GST (Goods and Services Tax)**: A comprehensive tax levied on the supply of goods and services in India, designed to replace multiple indirect taxes and streamline the tax process.
- **Machine Learning**: A branch of artificial intelligence that allows computers to learn from data and make predictions or decisions without being explicitly programmed.
- **Random Forest Classifier**: A machine learning algorithm that builds multiple decision trees and merges their results to improve accuracy and avoid overfitting.
- **Correlation Matrix**: A table showing the correlation coefficients between different variables, used to identify relationships between them.
- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: A performance measurement for classification problems that shows how well the model distinguishes between classes.
- **Confusion Matrix**: A table used to describe the performance of a classification model by comparing predicted and actual values.
- **SMOTE (Synthetic Minority Over-sampling Technique)**: A technique used to handle imbalanced datasets by generating synthetic examples of the minority class to balance the data.
- **Feature Importance**: A ranking of how important each feature in the dataset is for the model's decision-making process.