Developing a Predictive Model for GST Analytics

Team ID: GSTN_821

Solution developed by undergraduate students
of Netaji Subhas University of Technology.

Problem Statement

Imbalanced Target Class

The dataset has a majority of 0s, making class 1 harder to predict.

Missing data

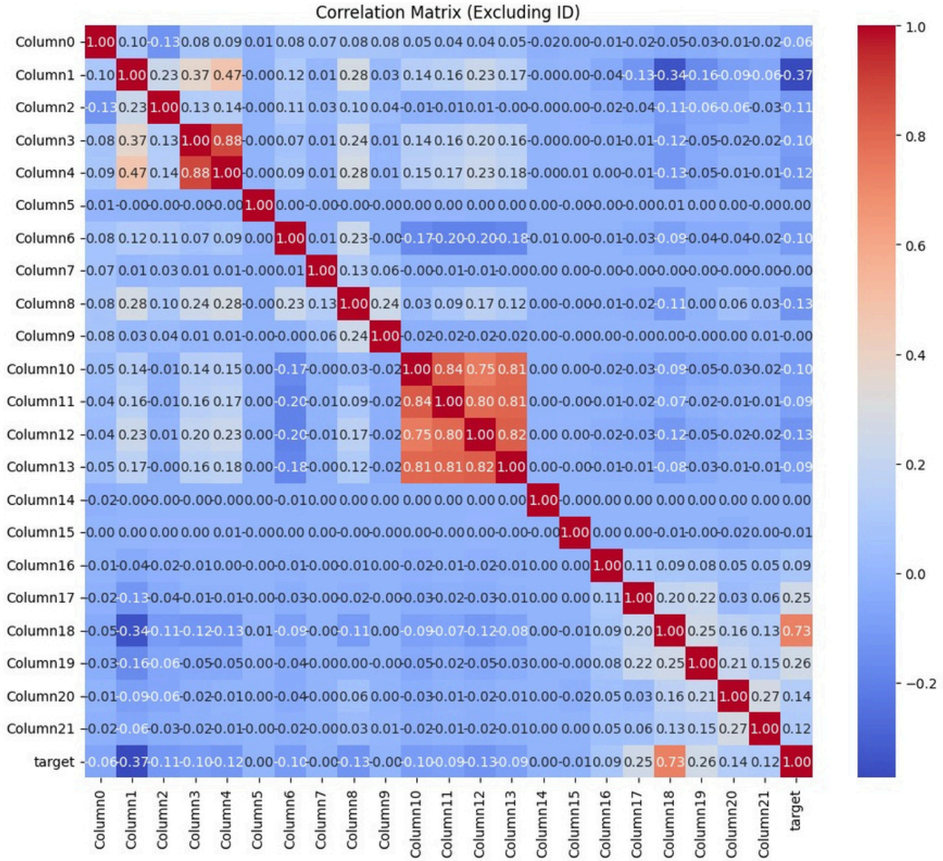
Some features had missing values, requiring careful imputation to ensure consistency.

Selecting Best Model

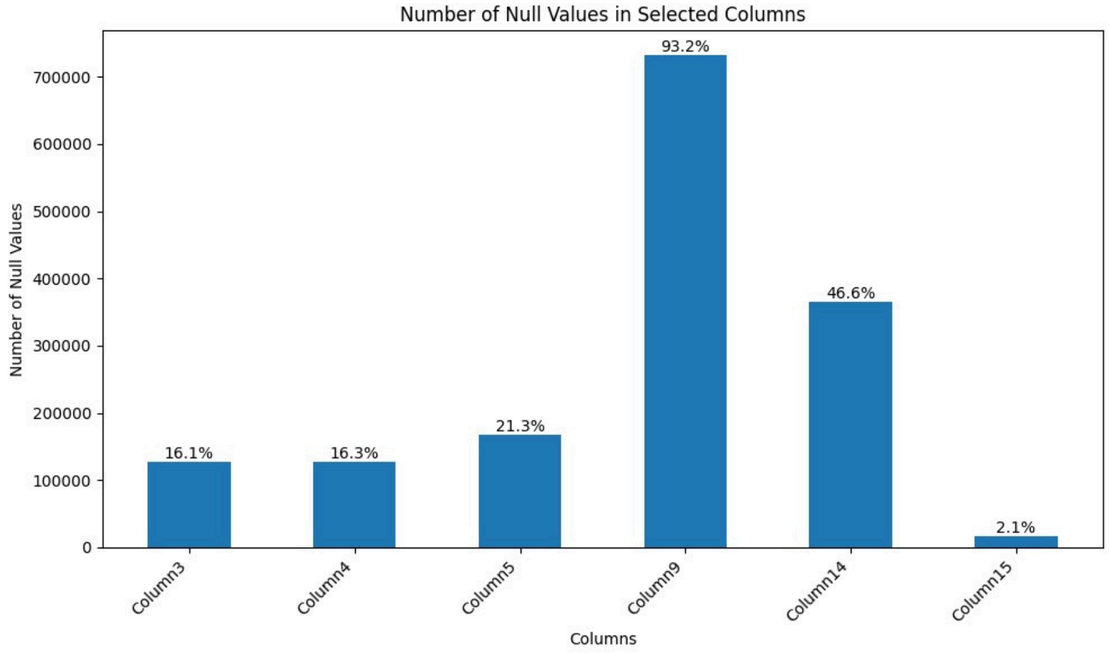
We evaluated models based on accuracy, efficiency, and their ability to handle large datasets.

Dataset Overview

X_train: 785133 records, ID+21 features
Y_train: 785133 records, ID+1 target(bool)



Correlation Matrix Heatmap: Displaying relationships between numeric features.



Null Values Bar Chart: Showing the proportion of missing data across features.

Data Processing

Median Imputation
In Column 5, 6, 9, 14, 15

Iterative Imputation
In column 3 and 4

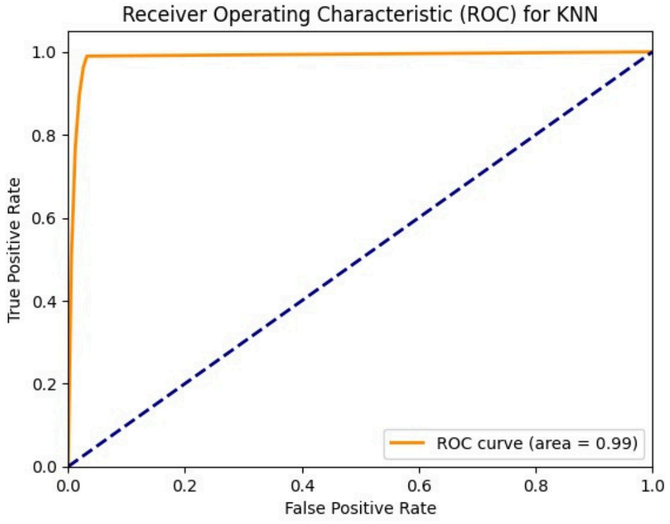
KNN Imputation
In Column 8 (k=5)

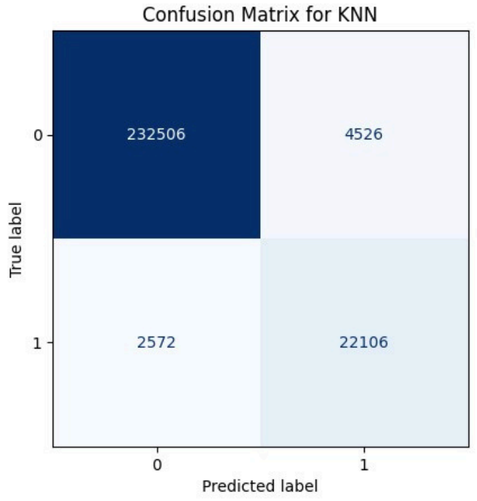
Standard Scaling
before training

Model Training & Performance

KNN

Accuracy: 0.9729
Precision: 0.8301
Recall: 0.8958
F1 Score: 0.8617
AUC-ROC: 0.9867
Confusion Matrix:
[[232506 4526]
[2572 22106]]
Log Loss: 0.2525
Balanced Accuracy: 0.9383





Catboost

Accuracy: 0.9783
Precision: 0.8485
Recall: 0.9369
F1 Score: 0.8905
AUC-ROC: 0.9947
Confusion Matrix:
[[232902 4130]
[1556 23122]]
Log Loss: 0.0501
Balanced Accuracy: 0.9598