GSTN Analytics Hackathon

Developing a Predictive Model in GST

Team ID: GSTN_435

Members:

- Mohit Gupta
- Harjot Singh
- Gaurav Kumar Chaurasiya
- Adamya Gaur

Exploratory Data Analysis

- On comprehensive analysis including the range, mean, and standard deviation of variables, we found **Column9** of data was missing for over **93%** of the feature.
- To look at correlations between features to see if there are some multicollinearity problems that could negatively affect the model, we used a heatmap.
- The Target variable suffers from a class imbalance effect since this problem is dominated by entries belonging to one class, which can affect our model's training.

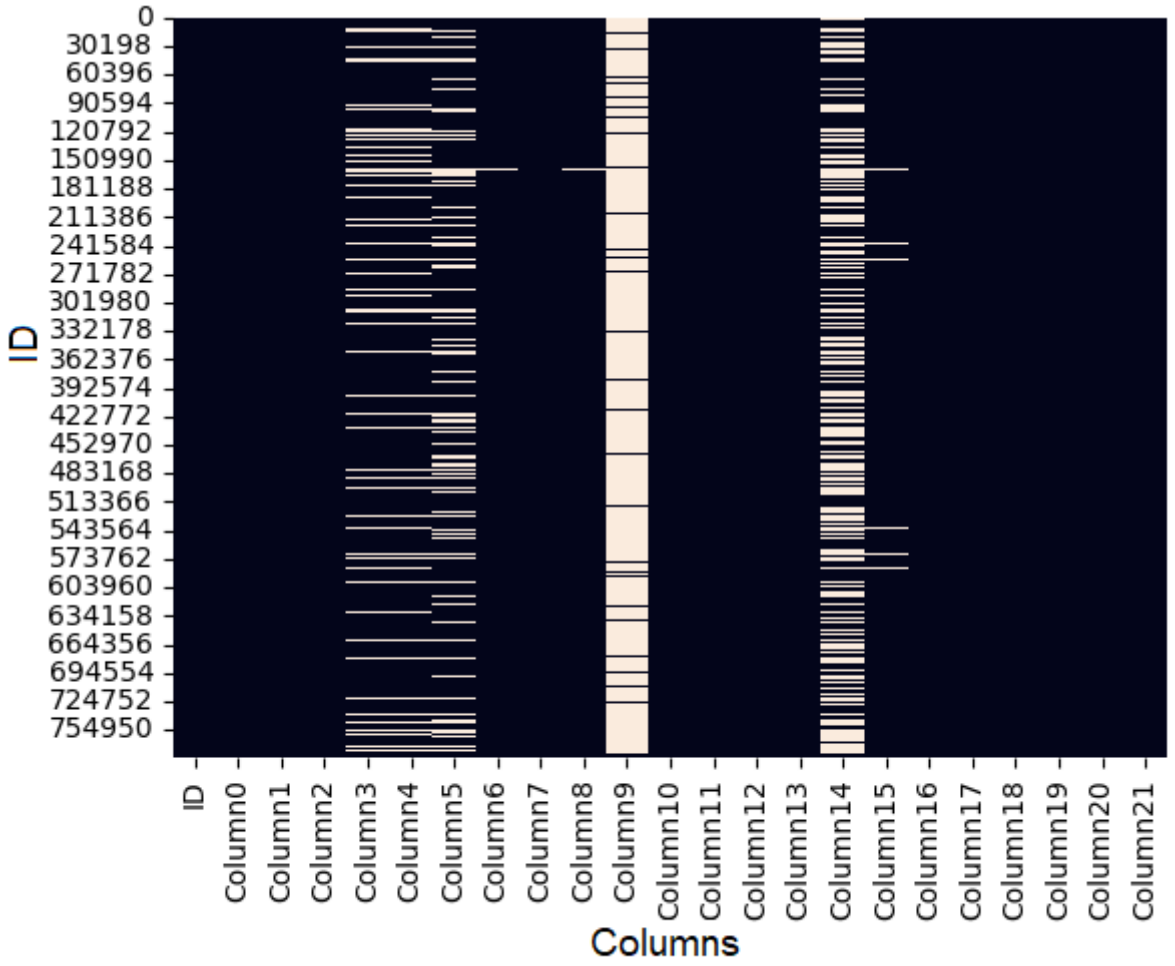


Fig: Missing values in columns

Data Preprocessing

Dropping Column

- Dropped Column9. It had more than 93% missing data.
- Imputing that much of missing data might result in bias and noise, so we decided to delete it completely.

Median Imputation

- Median is primarily outlier resistant and constitutes an effective value that does not get dominated by the extreme value as well.

Standard Scaling

- This step ensured that all variables were on a similar scale, facilitating better convergence during model training and improving the overall performance of our algorithms.

Data Splitting

- The data was split into training and validation sets to ensure the model's performance would generalize well to unseen data.
- 80% of the data was used for training and 20% for validation.

Model Selection

- We identified three algorithms that demonstrate strong performance and adaptability:

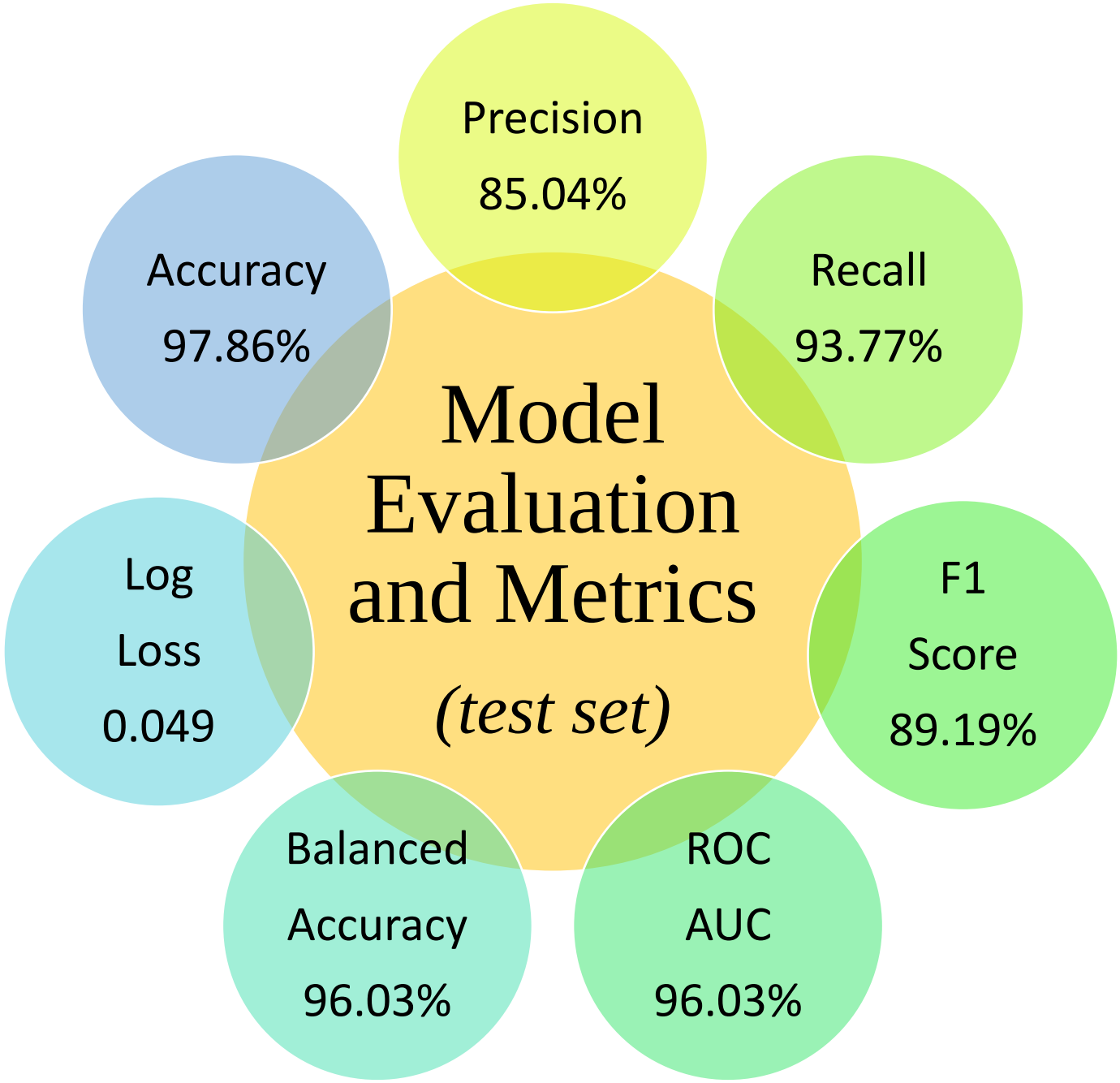
  - **XGBoost** is well-regarded for its robustness its regularization capabilities (L1 and L2) help reduce overfitting, making it a suitable choice.
  - **CatBoost** is particularly effective in managing features and offers strong performance even in the presence of missing values.
  - **LightGBM** is optimized for speed and efficiency, making it suitable for large datasets while maintaining competitive accuracy.

- To enhance our model performance further, we implemented a custom voting classifier that combines the predictions of these three models. The **CustomVotingClassifier** aggregates the outputs from individual models using **majority voting**.

- The custom voting mechanism not only combines the strengths of each individual model but also enhances overall predictive accuracy.

Hyperparameter Tuning

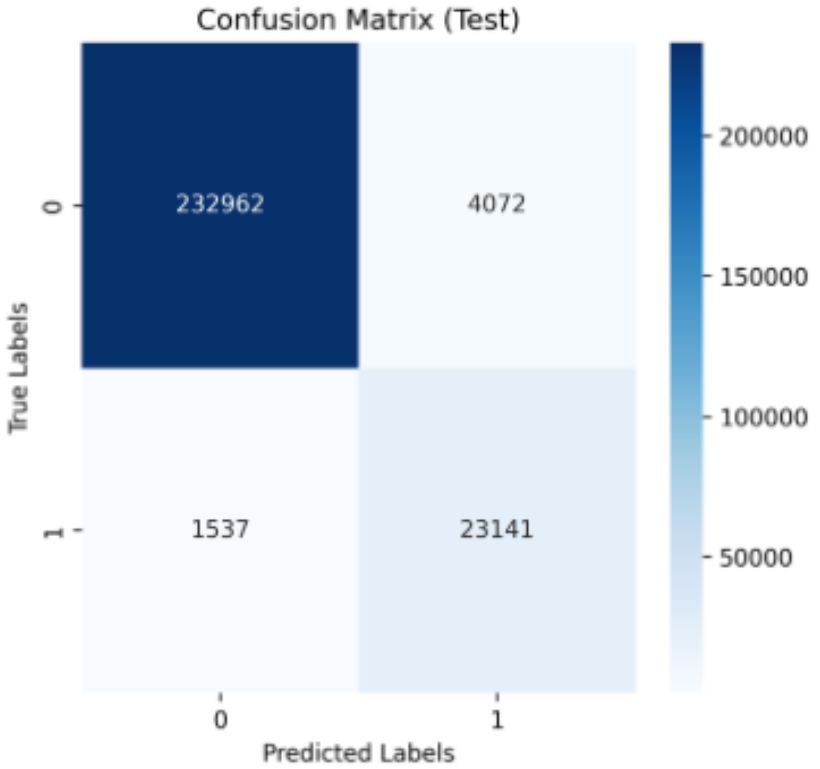
- Hyperparameter tuning using both GridSearchCV and Optuna.
- First, we have used GridSearchCV to search over various parameters including n_estimators, max_depth, learning_rate, subsample, and colsample_bytree.
- We placed Optuna on top of this to further enhance our tuning process. Optuna is a hyperparameter optimization framework that employs a more efficient approach by using TPE in sampling the hyperparameters.


Precision
85.04%
Accuracy
97.86%
Recall
93.77%
Model Evaluation and Metrics
(test set)
Log Loss
0.049
F1 Score
89.19%
Balanced Accuracy
96.03%
ROC AUC
96.03%

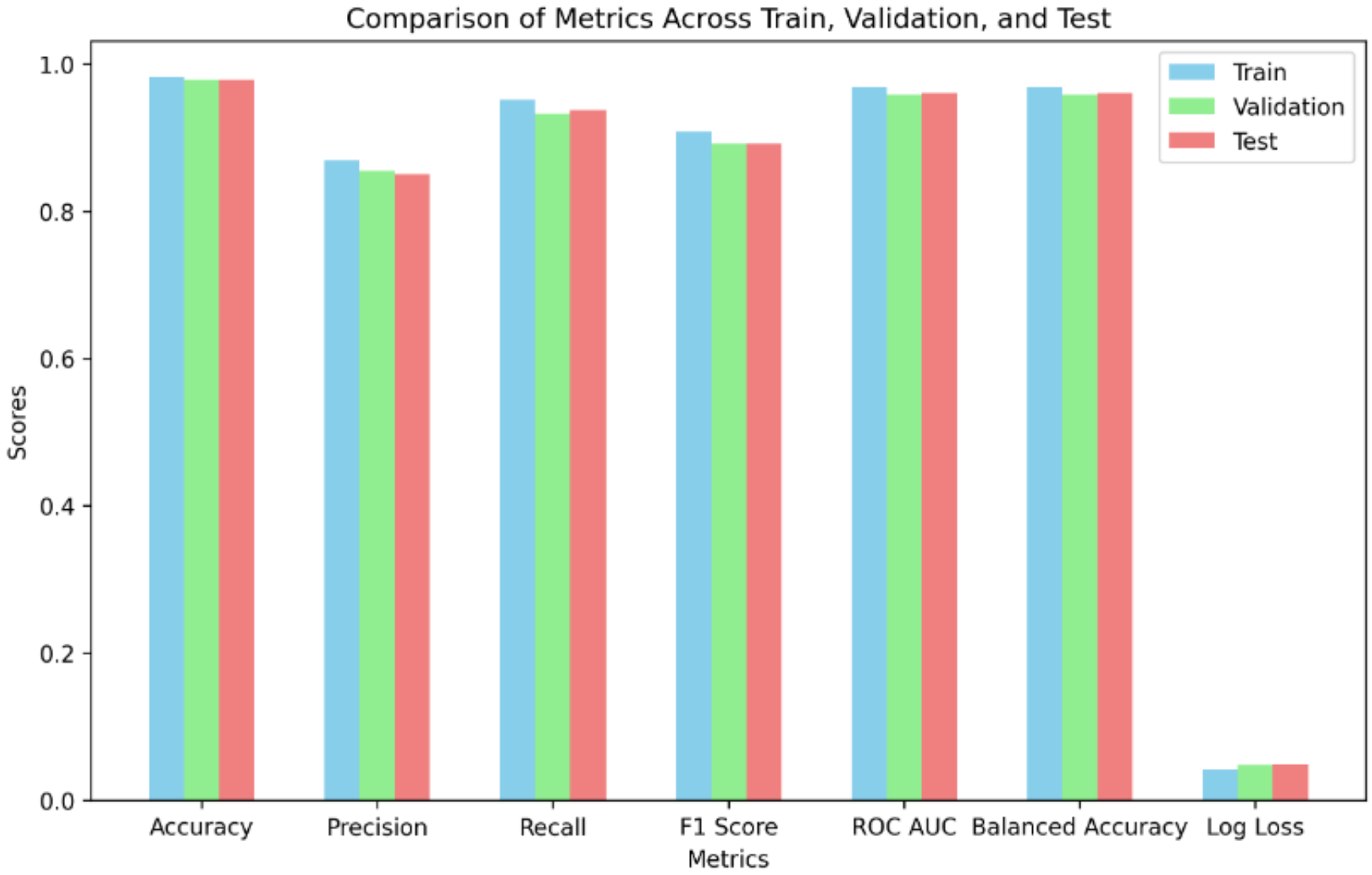
Results

The following results and confusion matrix were obtained:

Class	Metric	Precision	Recall	F1 Score	Support
Class 0	Training	0.99	0.99	0.99	568,880
	Validation	0.99	0.98	0.99	142,220
	Testing	0.99	0.98	0.99	237,034
Class 1	Training	0.87	0.95	0.91	59,226
	Validation	0.85	0.93	0.89	14,807
	Testing	0.85	0.94	0.89	24,678



Visual Comparison of metrics



Conclusion

- The model demonstrates high accuracy, precision, recall, and F1 scores across training, validation, and testing datasets, highlighting its effectiveness in handling a highly imbalanced classification problem.

- The metrics indicate strong performance for the minority class, suggesting that the model is well-tuned for practical applications. The consistent results across datasets also imply that the model is robust and generalizes well.

- Further improvements could focus on reducing false positives and enhancing interpretability for end-users.

- We believe our effective and insightful solution might strengthen the GST analytics framework and contribute to the nation’s progress.


Thankyou
for
your time