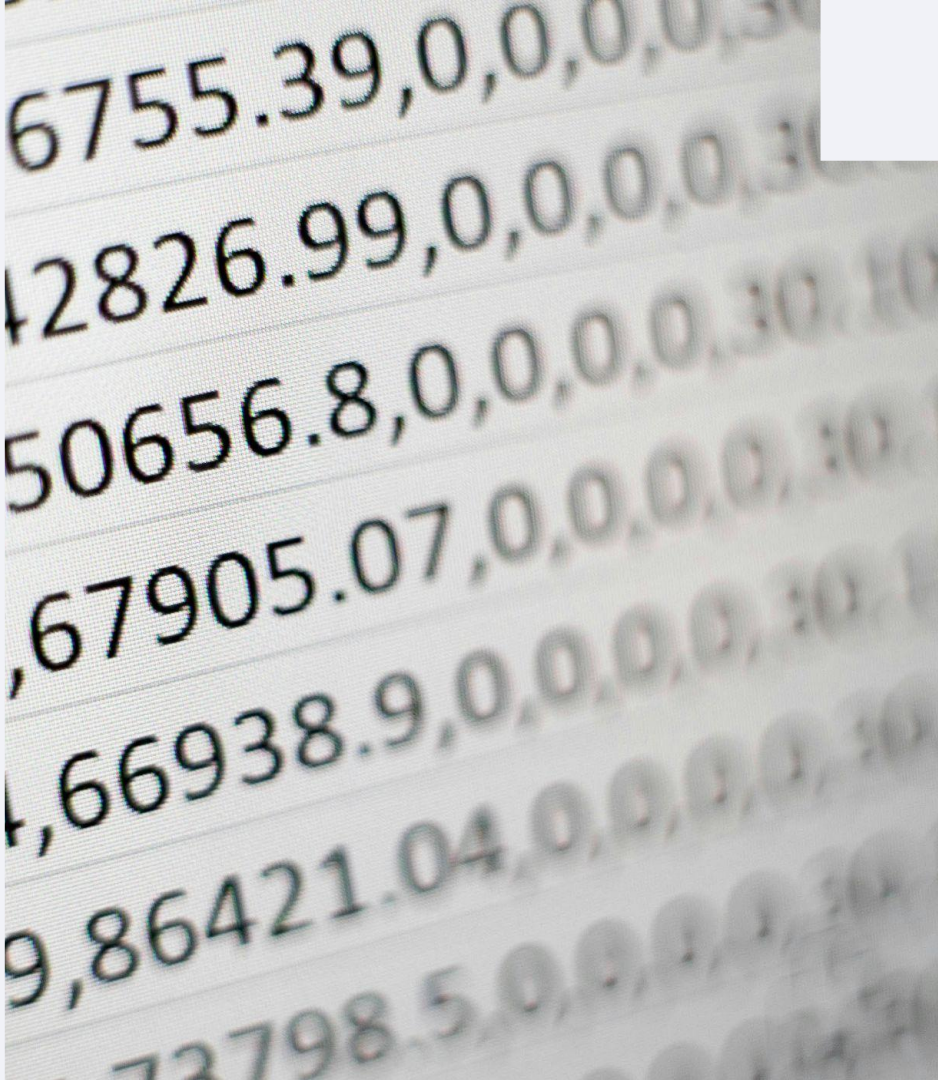

# Fraud Detection Project

Akash Chandra Baidya

# Dataset Overview

- Collected from Kaggle[1]
- Fraud transactions from 2019 - 2020
- Credit cards of 1000 customers doing transactions with 800 merchants
- 23 columns and 556k rows

QR Code to access project:



# Descriptive Analysis

- Utilized Power BI to visualize different aspects of the dataset in detail
- Randomly selected 4000 rows for training and testing
- Used different statistical methods to get a descriptive idea of the data
-

| Metric | Count | Mean | Std | Min | 25% | 50% | 75% | Max | Missing | Unique |
|---|---|---|---|---|---|---|---|---|---|---|
| Unnamed: 0 | 4000 | 277676.99 | 161805.1 | 53 | 138672.75 | 275161.5 | 420327.25 | 555650 | 0 | 4000 |
| cc_num | 4000 | 38933759606985 | 12704056209766 | 60416207185 | 18001145325019 | 35218152160915 | 45875771611606 | 49923463980651 | 0 | 857 |
| amt | 4000 | 66.36 | 126.63 | 1 | 9.59 | 46.02 | 82.07 | 5044.68 | 0 | 3278 |
| zip | 4000 | 48718.28 | 26657.14 | 1257 | 25442 | 48202 | 71880.5 | 99783 | 0 | 847 |
| lat | 4000 | 38.64 | 5.03 | 20.03 | 34.87 | 39.4 | 42.07 | 64.76 | 0 | 846 |
| long | 4000 | -90.06 | 13.54 | -165.67 | -96.7 | -87.36 | -80.13 | -67.95 | 0 | 846 |
| city_pop | 4000 | 97714.63 | 335644.03 | 23 | 741 | 2408 | 20639.75 | 2906700 | 0 | 779 |
| unix_time | 4000 | 1380661963.64 | 5235791.24 | 1371817839 | 1376021000.5 | 1380640566.5 | 1385919396.25 | 1388532811 | 0 | 4000 |
| merch_lat | 4000 | 38.65 | 5.04 | 19.45 | 34.97 | 39.45 | 41.98 | 65.37 | 0 | 4000 |
| merch_long | 4000 | -90.07 | 13.55 | -166.54 | -96.8 | -87.38 | -80.11 | -67.05 | 0 | 4000 |
| is_fraud | 4000 | 0 | 0.05 | 0 | 0 | 0 | 0 | 1 | 0 | 2 |

# Data Preprocessing and Statistics

Before analysis, the data underwent preprocessing:

- **Removed irrelevant columns** like 'Unnamed: 0', 'street', etc.
- Applied **one-hot encoding** for categorical features.
- **Standardized numerical features** for model compatibility.

# Imbalanced Dataset

Dataset contains more valid transactions than fraud cases !

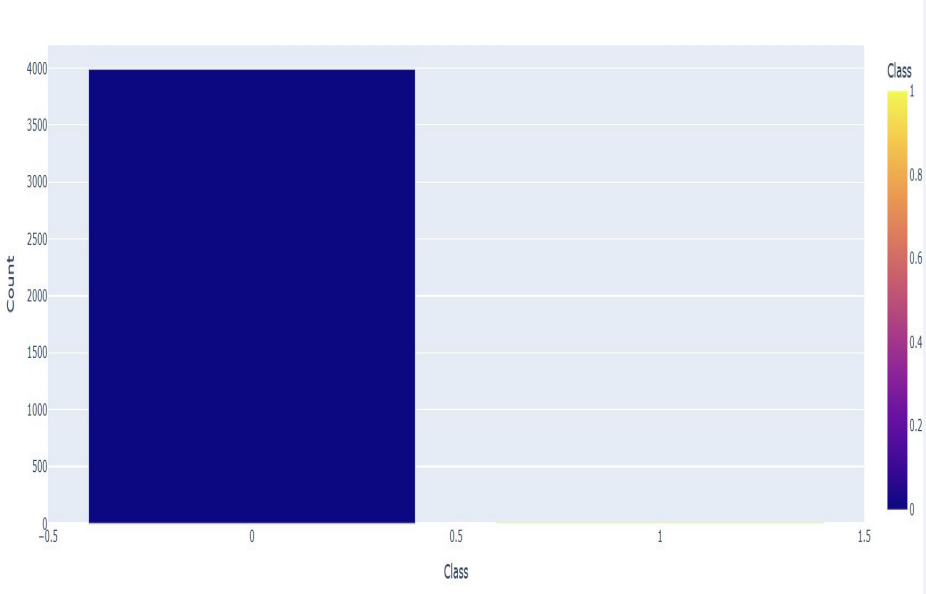



Fig: Imbalanced dataset overview

# Class Imbalance Resolution

The dataset was highly imbalanced, with only 0.2% of records labeled as 'Fraud'.
To address this:

- Applied **SMOTE (Synthetic Minority Oversampling Technique)**.
- Balanced the classes to ensure **fair model training and evaluation**.

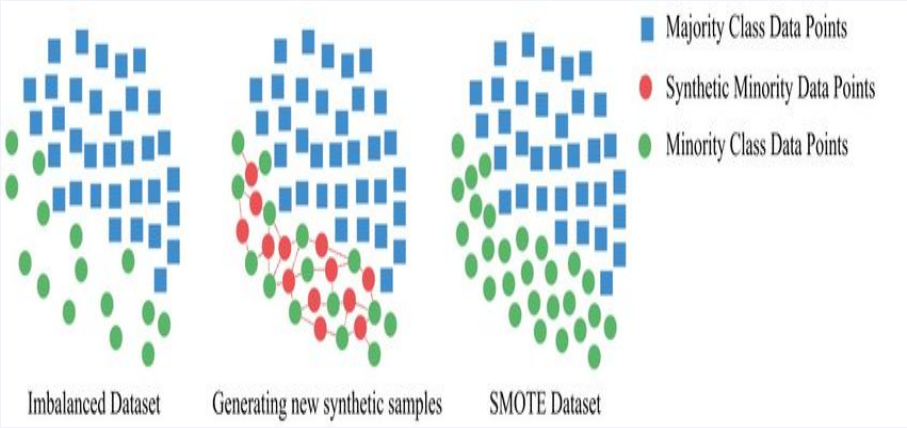



Fig: SMOTE

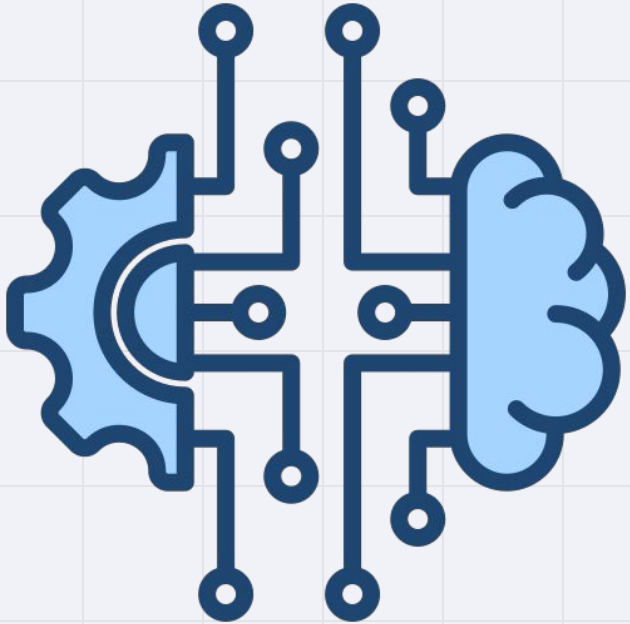

Fig: Balanced dataset overview after SMOTE

# Cross-Validation and Data Leakage Prevention

To ensure reliable and unbiased results:

- **Stratified K-Fold Cross-Validation** was used for consistent class representation across folds.
- Data leakage was prevented by ensuring feature **standardization occurred independently for training and testing sets**.

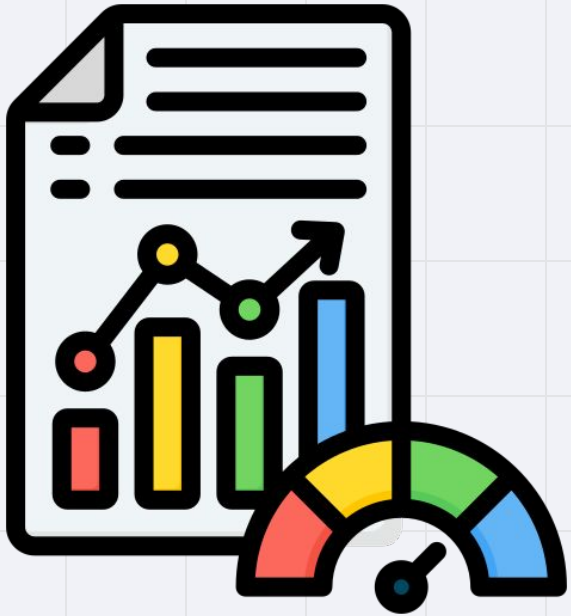
# Models Applied



Three machine learning models were used:

- **Logistic Regression**: Simple and interpretable.
- **Random Forest**: Ensemble method leveraging multiple decision trees.
- **Gradient Boosting**: Focused on minimizing prediction errors iteratively.

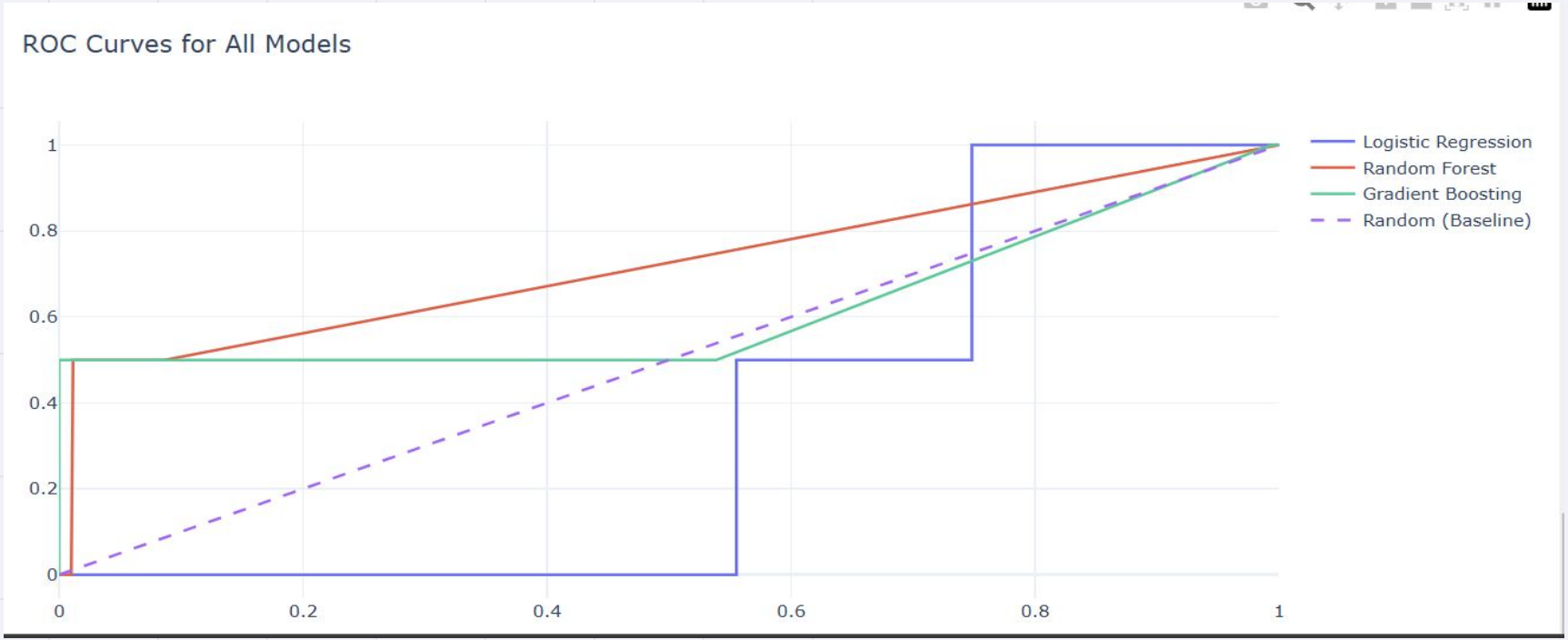
# Evaluation Metrics



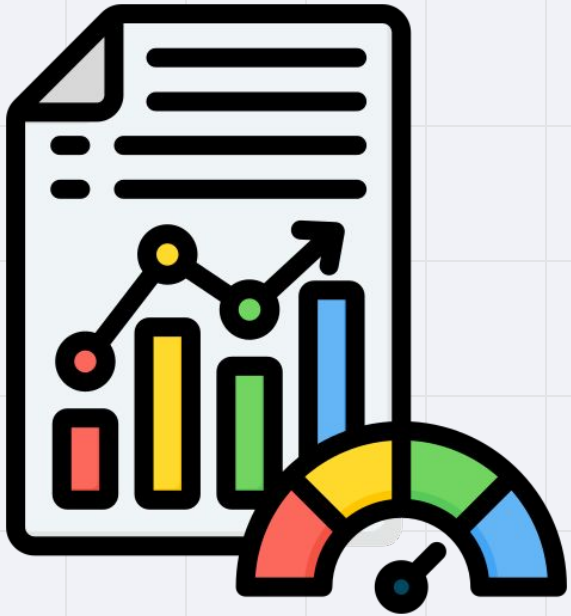
Models were evaluated using:

- **Precision:** Proportion of true frauds among predicted frauds.
- **Recall:** Proportion of actual frauds correctly identified.
- **F1-Score:** Harmonic mean of precision and recall.
- **AUC-ROC:** Measure of model's ability to distinguish classes.

# ROC CURVE


ROC Curves for All Models
Logistic Regression
Random Forest
Gradient Boosting
Random (Baseline)
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

# Results and Comparison



**Key Findings:**

- **Logistic Regression**: Balanced across metrics, good generalization.
- **Random Forest:** High precision but **lower recall**.
- **Gradient Boosting:** Best overall performance with **high recall and AUC-ROC**.

# Interactive Dashboard


A dynamic dashboard was built to:

- Visualize model metrics like precision, recall, and accuracy.
- Display confusion matrices and ROC curves.
- Enable result export in multiple formats (CSV, JSON, Excel).

The dashboard provides actionable insights for fraud detection improvement.

# Thank you !