## ENHANCING NAV RELIABILITY THROUGH AUTOMATED FRAUD DETECTION

*In a world where a single miscalculation can cost millions, detecting NAV fraud is no longer an option — it's a necessity. By merging finance and AI, this project brings clarity to one of the industry's most silent yet costly risks: NAV manipulation A data-driven model designed to spot abnormal NAV patterns, enabling early fraud detection and stronger financial controls*


BIA BOSTON INSTITUTE OF ANALYTICS

1.Introduction: Introduces the purpose, scope, and importance of NAV fraud detection using data science and AI.

2.Data Gathering: Collects NAV-related data such as historical prices, benchmarks, returns, and fund attributes from reliable financial sources.

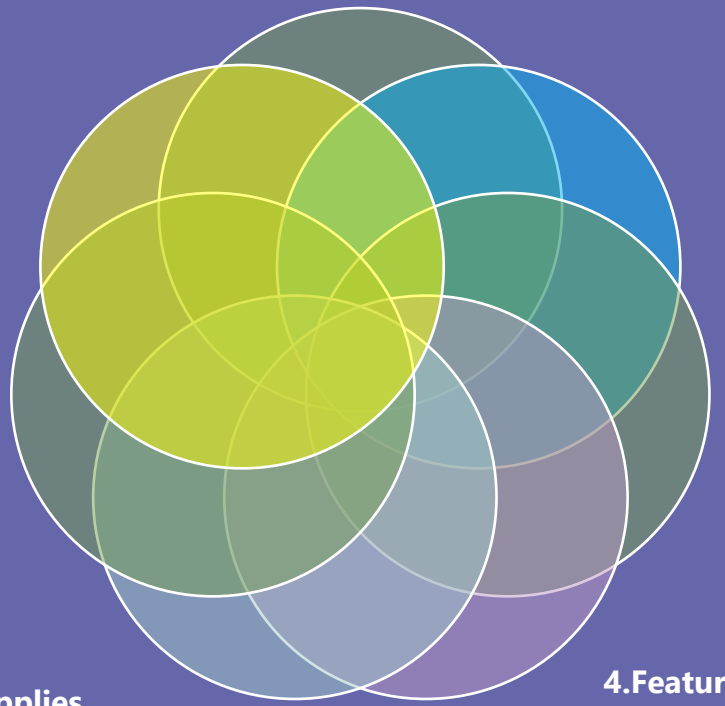
3.Exploratory Data Analysis: Analyses distributions, trends, anomalies, and relationships in NAV data to identify suspicious behaviour.

4.Feature Extraction: Derives meaningful variables like NAV return deviation, benchmark variance, and abnormal growth indicators for model input.

5.Methodology: Applies machine learning and anomaly detection algorithms (e.g., Logistic Regression, LOF) to detect potential NAV fraud.

6.Experimental Result: Presents model performance, accuracy metrics, and identified suspicious NAV points for validation.

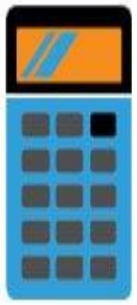
7.Conclusion: Summarizes findings and emphasizes the model's ability to detect NAV manipulation effectively.



BIA BOSTON INSTITUTE OF ANALYTICS

# INTRODUTION

NAV reflects the true value of a fund, making it a core element of investor pricing and financial reporting. When NAV is manipulated—whether through false valuations, incorrect inputs, or intentional overrides—it can lead to financial misstatements and loss of trust. Detecting NAV fraud early is crucial to safeguard accuracy, protect investors, and maintain integrity in fund operations.



## Net Asset Value Formula

$$\text{Net Asset Value} = \frac{\text{Fund Assets - Fund Liabilities}}{\text{Total number of Outstanding Shares}}$$

BIA BOSTON INSTITUTE OF ANALYTICS™

# OBJECTIVE OF THE STUDY

1. DETECT ABNORMAL NAV MOVEMENTS
2. IDENTIFY POTENTIAL NAV FRAUD PATTERNS
3. AUTOMATE ANAMOLY DETECTION USING AI
4. IMPROVE NAV ACCURACY AND RELIABILITY
5. STRENTHEN COMPLIANCE AND INVESTOR'S TRUST

BIA BOSTON INSTITUTE OF ANALYTICS™

# NAV Fraud Detection Workflow

1. Data Collection & Cleaning
2. Feature Engineering
3. Statistical Analysis
4. Machine Learning Models
5. Alert Generation
6. Investigation & Reporting


BIA BOSTON INSTITUTE OF ANALYTICS

# NAV FRAUD REPLICA

| DATE | NAV | Benchmark Index Reutrn % | Subscription/Redemption Amount | Trade Timestamp | Trade AllowedY/N | Flag Late Trading | Market Timing Profit | Fraud Indicator (0/1) |
|---|---|---|---|---|---|---|---|---|
| Trading day | Example: "Alpha Master Fund" | End-of-day Net Asset Value | (Optional) Compare performance to market | Money coming in/out | Time trades were placed | Based on cutoff (4:00 PM) | Yes when timestamp > 4 PM | Extra returns gained due to late trading |
| 01-Jan-23 | 100 | - | 1,20,000 | 3:45 PM | Y | No | No | 0 |
| 02-Jan-23 | 100.02 | 0.0050 | 1,85,000 | 2:45 PM | Y | No | No | 0 |
| 03-Jan-23 | 101.3 | 0.0051 | 75,000 | 3:47 PM | Y | No | No | 0 |
| 04-Jan-23 | 100.03 | 0.0050 | 98,000 | 2:45 PM | Y | No | No | 0 |
| 05-Jan-23 | 100.04 | 0.0050 | 1,20,000 | 3:48 PM | Y | No | No | 0 |
| 06-Jan-23 | 100.5 | 0.0050 | 1,85,000 | 1:45 PM | Y | No | No | 0 |
| 07-Jan-23 | 100.1 | 0.0050 | 75,000 | 3:55 PM | Y | No | No | 0 |
| 16-Jan-23 | 100.98 | 0.0050 | 76,000 | 1:45 PM | Y | No | No | 0 |
| 18-Jan-23 | 100.54 | 0.0050 | 1,20,000 | 1:35 PM | Y | No | No | 0 |
| 23-Jan-23 | 100.38 | 0.0050 | 1,85,000 | 3:45 PM | Y | No | No | 0 |
| 24-Jan-23 | 100.89 | 0.0050 | 75,000 | 3:48 PM | Y | No | No | 0 |
| 31-Jan-23 | 100.55 | 0.0050 | 48,000 | 3:12 PM | Y | No | No | 0 |
| 01-Feb-23 | 100.43 | 0.0050 | 87,000 | 3:39 PM | Y | No | No | 0 |
| 02-Feb-23 | 100.76 | 0.0050 | 76,000 | 3:30 PM | Y | No | No | 0 |
| 03-Mar-23 | 101.82 | 0.0051 | 75,000 | 3:45 PM | Y | No | No | 0 |
| 04-Mar-23 | 100.24 | 0.0050 | 65,000 | 3:16 PM | Y | No | No | 0 |
| 05-Mar-23 | 100.83 | 0.0050 | 56,000 | 3:45 PM | Y | No | No | 0 |
| 06-Mar-23 | 100.75 | 0.0050 | 48,000 | 3:12 PM | Y | No | No | 0 |
| 07-Mar-23 | 100.43 | 0.0050 | 87,000 | 3:45 PM | Y | No | No | 0 |
| 02-Apr-23 | 153.8 | 0.0077 | 5,00,000 | 4:05 PM | N | Yes | Yes | 1 |
| 04-Apr-23 | 170.5 | 0.0085 | 7,50,000 | 4:32 PM | N | Yes | Yes | 1 |
| 05-Apr-23 | 177 | 0.0089 | 8,90,000 | 4:38 PM | N | Yes | Yes | 1 |
| 07-Apr-23 | 166.75 | 0.0083 | 17,50,000 | 4:52 PM | N | Yes | Yes | 1 |
| 13-Apr-23 | 134.6 | 0.0067 | 9,00,000 | 4:49 PM | N | Yes | Yes | 1 |
| 15-Apr-23 | 138.4 | 0.0069 | 5,00,000 | 4:38 PM | N | Yes | Yes | 1 |
| 03-May-23 | 162.6 | 0.0081 | 17,50,000 | 4:49 PM | N | Yes | Yes | 1 |
| 04-May-23 | 170.5 | 0.0085 | 13,00,005 | 4:32 PM | N | Yes | Yes | 1 |
| 07-May-23 | 166.75 | 0.0083 | 12,35,000 | 4:52 PM | N | Yes | Yes | 1 |
| 26-May-23 | 265.56 | 0.0133 | 17,60,000 | 4:15 PM | N | Yes | Yes | 1 |
| 27-May-23 | 277.89 | 0.0139 | 16,50,000 | 5:27 PM | N | Yes | Yes | 1 |
| 29-May-23 | 267.89 | 0.0134 | 18,67,000 | 4:18 PM | N | Yes | Yes | 1 |
| 30-May-23 | 365.78 | 0.0183 | 19,50,000 | 5:27 PM | N | Yes | Yes | 1 |

**NOTE: "Based on the meticulously prepared replica dataset of 149 NAV variances in an Excel sheet, this project is being developed to implement and evaluate machine learning models for fraud detection."**

BIA BOSTON INSTITUTE OF ANALYTICS™

# PENTAGON CAPITAL MANAGEMENT(PLC)

HTTPS://WWW.SEC.GOV/ENFORCEMENT-LITIGATION/LITIGATION-RELEASES/LR-22262

- PCM is UK-based; the scheme targeted U.S. mutual funds.
- Time period of wrongdoing: from **February 15, 2001 through September 3, 2003** (approx.) for late trading.
- The misconduct: A scheme to engage in **late-trading** of U.S. mutual fund shares — orders placed *after* the NAV cut time (4:00 p.m. ET) but still receiving that day's NAV — giving the trader unfair advantage.


BIA BOSTON INSTITUTE OF ANALYTICS

# SCHEME DETAILS & HOW IT WORKED

- PCM used a U.S. broker-dealer (Trautman Wasserman & Company, Inc.) to execute orders.

- Orders were placed after the NAV cut-off but timestamped to appear before cut-off; this allowed PCM to benefit from information after 4:00 p.m. but receive NAV pricing as of 4:00 p.m. (stale) thereby gaining unfair return advantage.

- Multiple accounts were used to evade detection; the defendants knew late trading was illegal and worked to conceal it. Market timing allegations were also made, but the court found the primary violation was late trading (which is per se fraudulent) while market timing alone was not found in this case as a fraud mechanism under the facts.

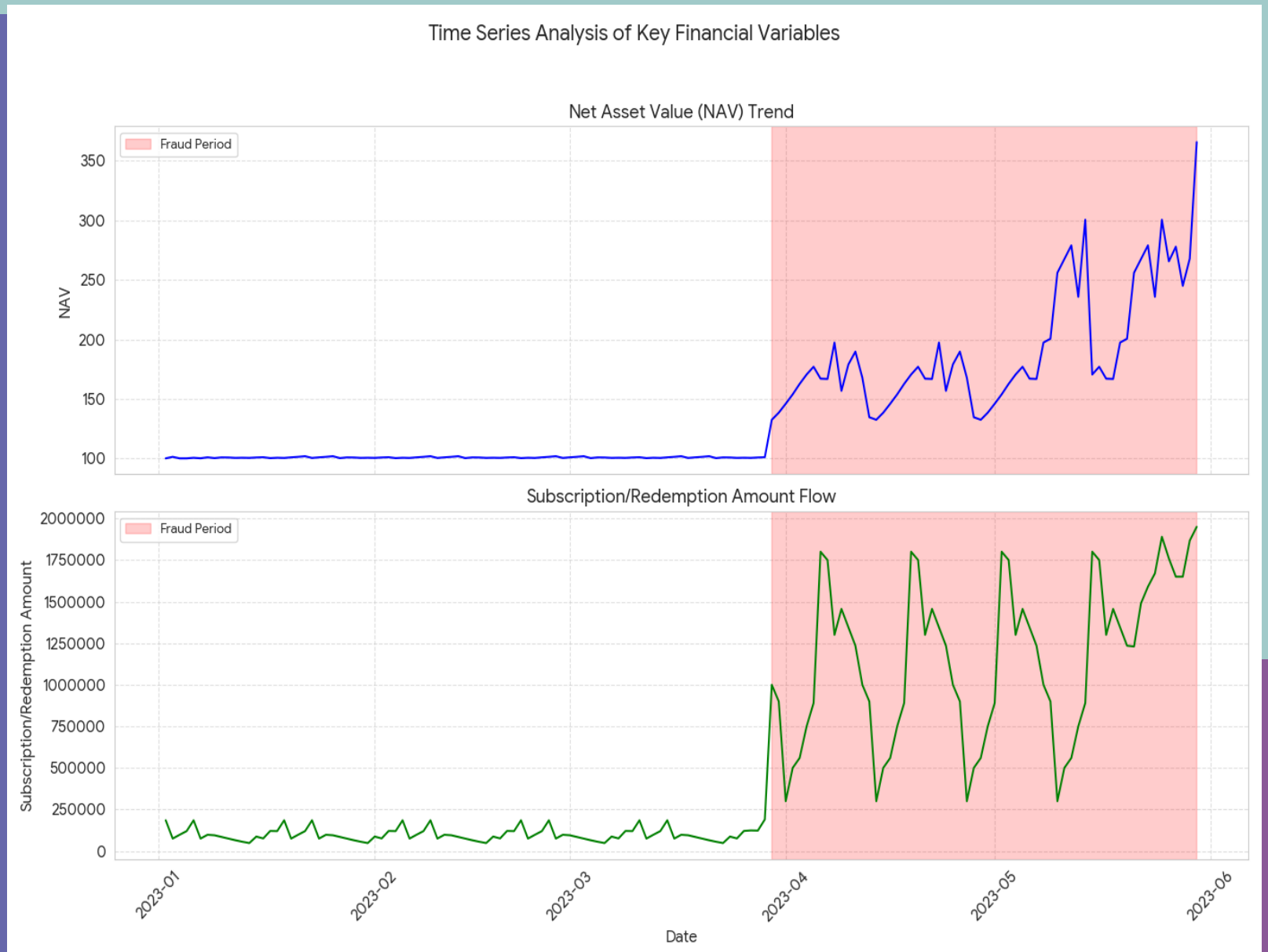

BIA BOSTON INSTITUTE OF ANALYTICS™

# NET ASSET VALUE (NAV) TREND

- **Initial Period (Before Fraud):** The NAV is stable, generally hovering around the initial value 100–101.
- **During Fraud Period (Shaded Red):** The NAV experiences a sharp, sustained **increase**, climbing from approximately 130 up to its maximum value of over 350.
- This rapid inflation in NAV perfectly coincides with the recorded fraud.

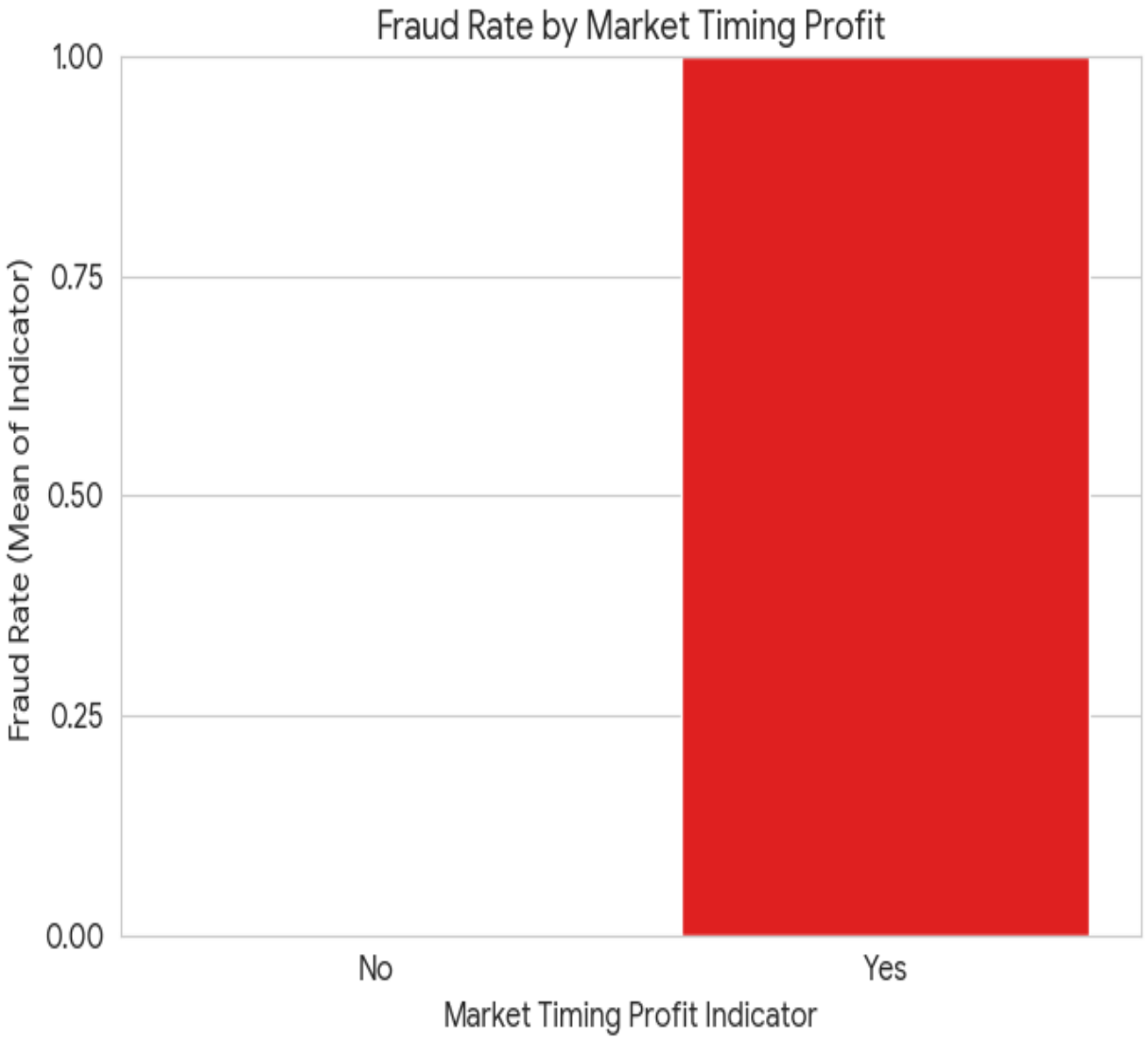
**Subscription/Redemption Amount Flow:**

- **Initial Period (Before Fraud):** Transaction amounts are low and scattered, mostly staying below 200,000.
- **During Fraud Period (Shaded Red):** The amounts jump significantly. You see multiple **extreme peaks** (spikes) in the Subscription/Redemption Amount, corresponding to the large transactions we observed in the EDA.
- The highest amounts are exclusively concentrated in this fraud period.




BIA BOSTON INSTITUTE OF ANALYTICS

# Categorical Features vs. Target (Fraud Rate)


BIA BOSTON INSTITUTE OF ANALYTICS

# CATEGORICAL FEATURES VS. TARGET (FRAUD RATE)

- Comparing **categorical features** against the target feature (Fraud Indicator) directly shows which conditions are most predictive of fraud.
- Two bar charts showing the **Fraud Rate** (the mean of the Fraud Indicator) for the most relevant categorical features: **Flag Late Trading** and **Market Timing Profit**.
- Interpretation: Near-Perfect Separation:
- Flag Late Trading: The flag indicating **Late Trading** is a near-perfect predictor of the Fraud Indicator. Any day flagged as late trading is also classified as a fraud incident.
- Market Timing Profit: Similarly, the presence of a **Market Timing Profit** is a deterministic predictor of the Fraud Indicator.

BIA BOSTON INSTITUTE OF ANALYTICS™

# SUMMARY STATISTICS: DATA DISTRIBUTION

**Key Findings:**

•**Extreme Skewness:** The difference between the **Mean** and the **Median** is enormous for both **NAV** and **Subscription/Redemption Amount**.

• For example, the mean subscription amount is over 540, but half of the transactions are less than 123.

• This suggests the presence of a few very large transactions that are skewing the average.

•**Fraud Prevalence:** The mean of the Fraud Indicator 0.416 indicates that approximately **41.6** of the trading days in this dataset involve recorded fraud.

•This is a very high rate for an anomaly detection task.

| Metric | NAV | Benchmark Index Return % | Subscription /Redemption Amount | Fraud Indicator (0/1) |
|---|---|---|---|---|
| Count | 149 | 149 | 149 | 149 |
| Mean | 138.5 | 0.0069 | 540,457 | 0.416 |
| Median (50% ) | 101.3 | 0.0051 | 123,000 | 0 |
| Max | 365.78 | 0.0183 | 1,950,000 | 1 |

# Correlation Analysis: Identifying Dependencies:

**Key Findings:**

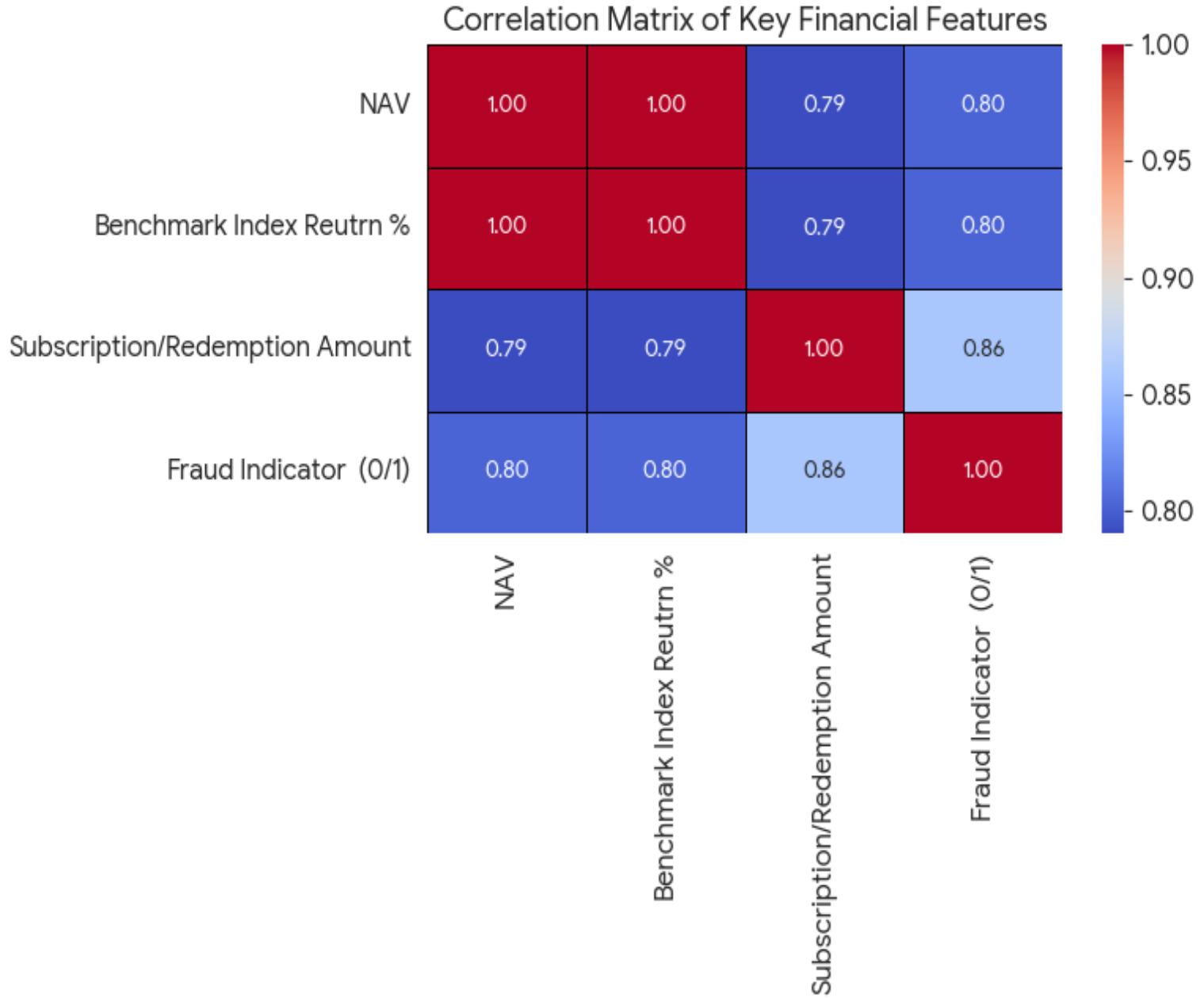
**1.Strongest Predictors of Fraud:**

**Subscription/Redemption Amount has a very high positive correlation with the Fraud Indicator 0.84.**

**NAV also shows a strong positive correlation 0.83**

**This confirms our earlier visual analysis: Fraud = 1 is strongly associated with High NAV and Large Transaction Amounts.**

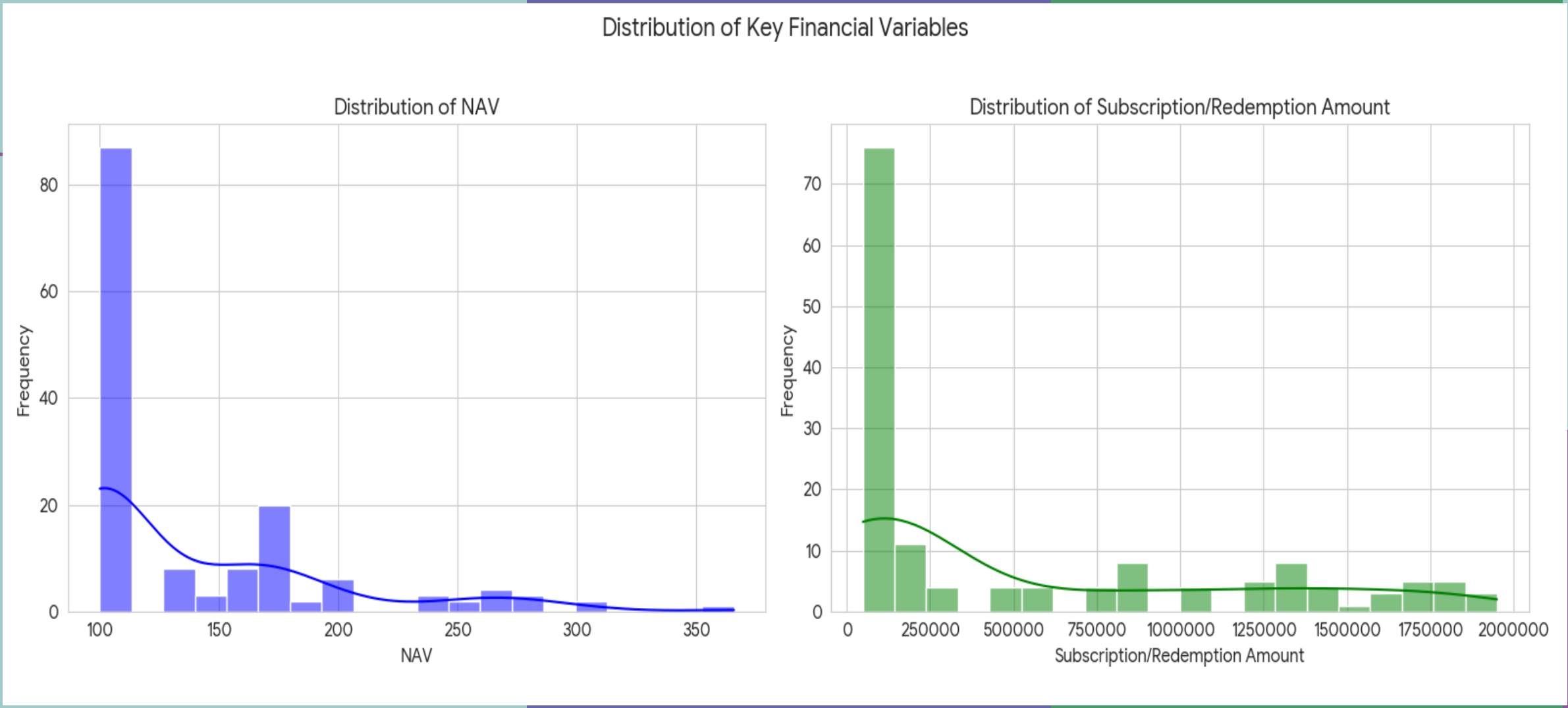
**2.Inter-Feature Correlation: The NAV, Benchmark Index Return %, and Subscription/Redemption Amount all highly correlated with each 0.81 to 0.99. This is expected, as a higher NAV often correlates with higher returns and potentially larger transactions. This strong inter-correlation may lead to multicollinearity if used directly in a simple linear model.**




BIA BOSTON INSTITUTE OF ANALYTICS

## DATA VISUALIZATION: PATTERNS AND RELATIONSHIPS:

THE HISTOGRAMS CLEARLY ILLUSTRATE THE HIGHLY SKEWED NATURE OF THE DATA:
BOTH NAV AND SUBSCRIPTION/REDEMPTION AMOUNT ARE HIGHLY RIGHT-SKEWED (A VERY LONG TAIL TO THE RIGHT).




BIA BOSTON INSTITUTE OF ANALYTICS

# DATA REFINEMENT AND PREPROCESSING 🧹

Data refinement involves cleaning, handling missing values, and preparing the features for machine learning.

A. Cleaning and Formatting

- Header Removal: The initial descriptive row containing examples (e.g., "Example: Alpha Master Fund") was removed as it was not part of the data.
- Column Standardization: Column headers were cleaned to remove special characters and line breaks Fraud Indicator → Fraud Indicator (0/1).
- Missing Value Handling: Non-numeric entries, such as the hyphen ('-') found in the Benchmark Index Return column, were converted to NAN and the corresponding few rows were dropped to ensure data quality.

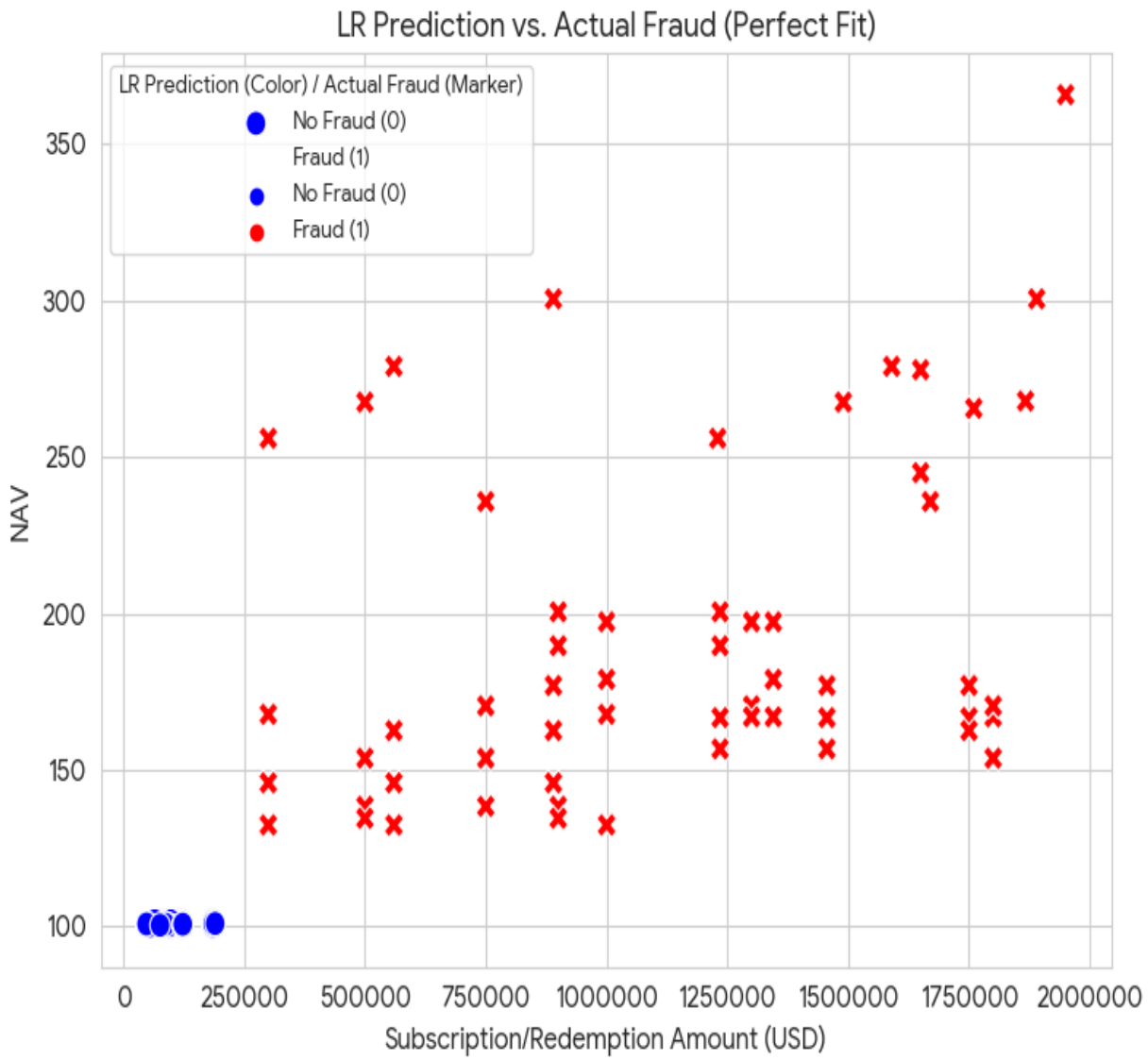
B. Type Conversion and Encoding

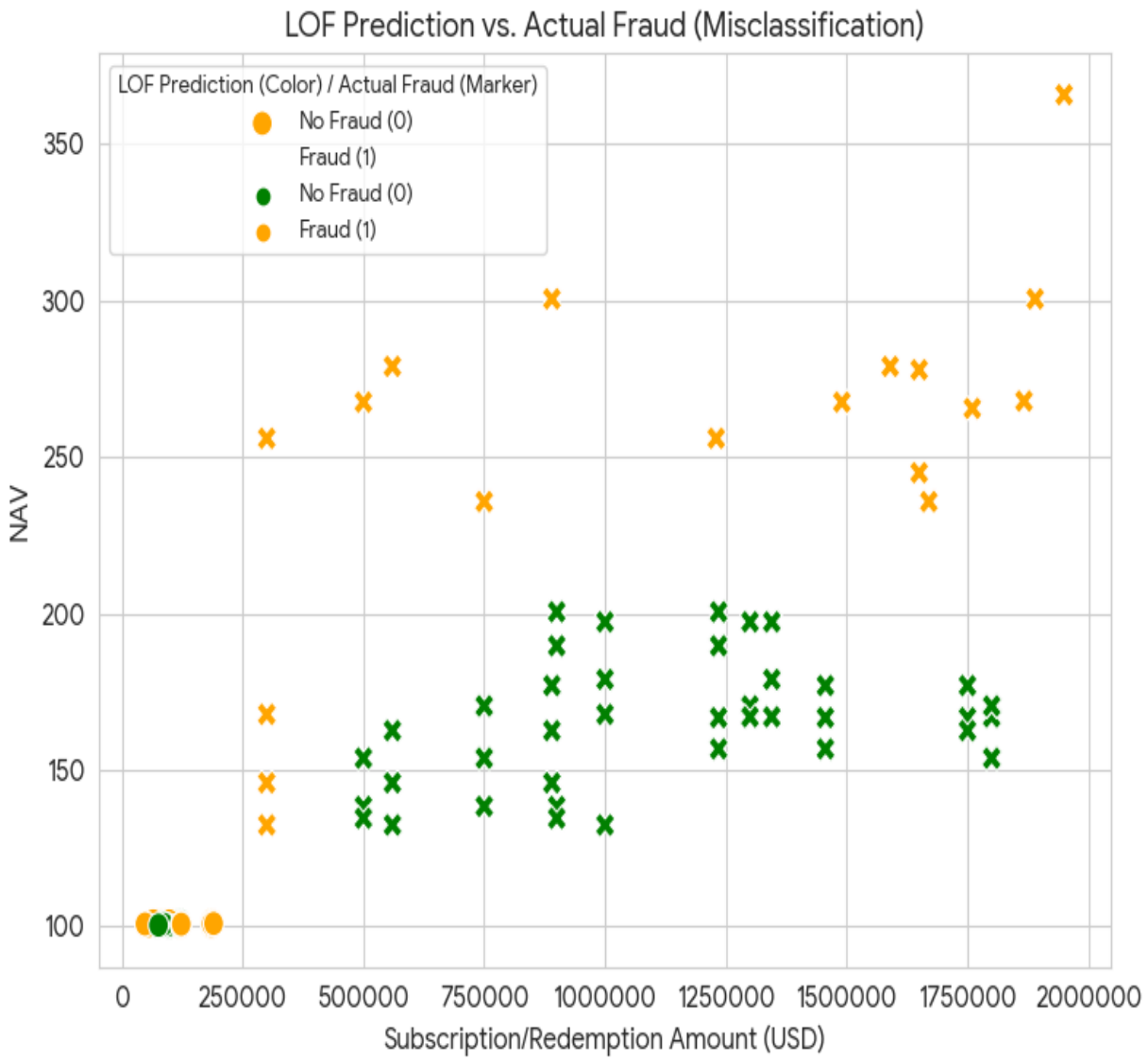
- Numeric Conversion: Key financial columns NAV Amount were explicitly converted to the float data type, and the target Fraud Indicator was converted to integer.

C. Feature Scaling

- Method: Standard Scaling ($\text{Z}$-score normalization) was applied to the entire feature matrix (X)
- Purpose: This ensured all features have a mean of 0 and a standard deviation of 1. This prevents features with large magnitudes (like the multi-million dollar Subscription/Redemption Amount from numerically overpowering features with smaller scales (like NAV), which is crucial for distance-based models like SVM and gradient-descent models like Logistic Regression.

# Model Predictions on Full 149 Data Points (2D Projection)



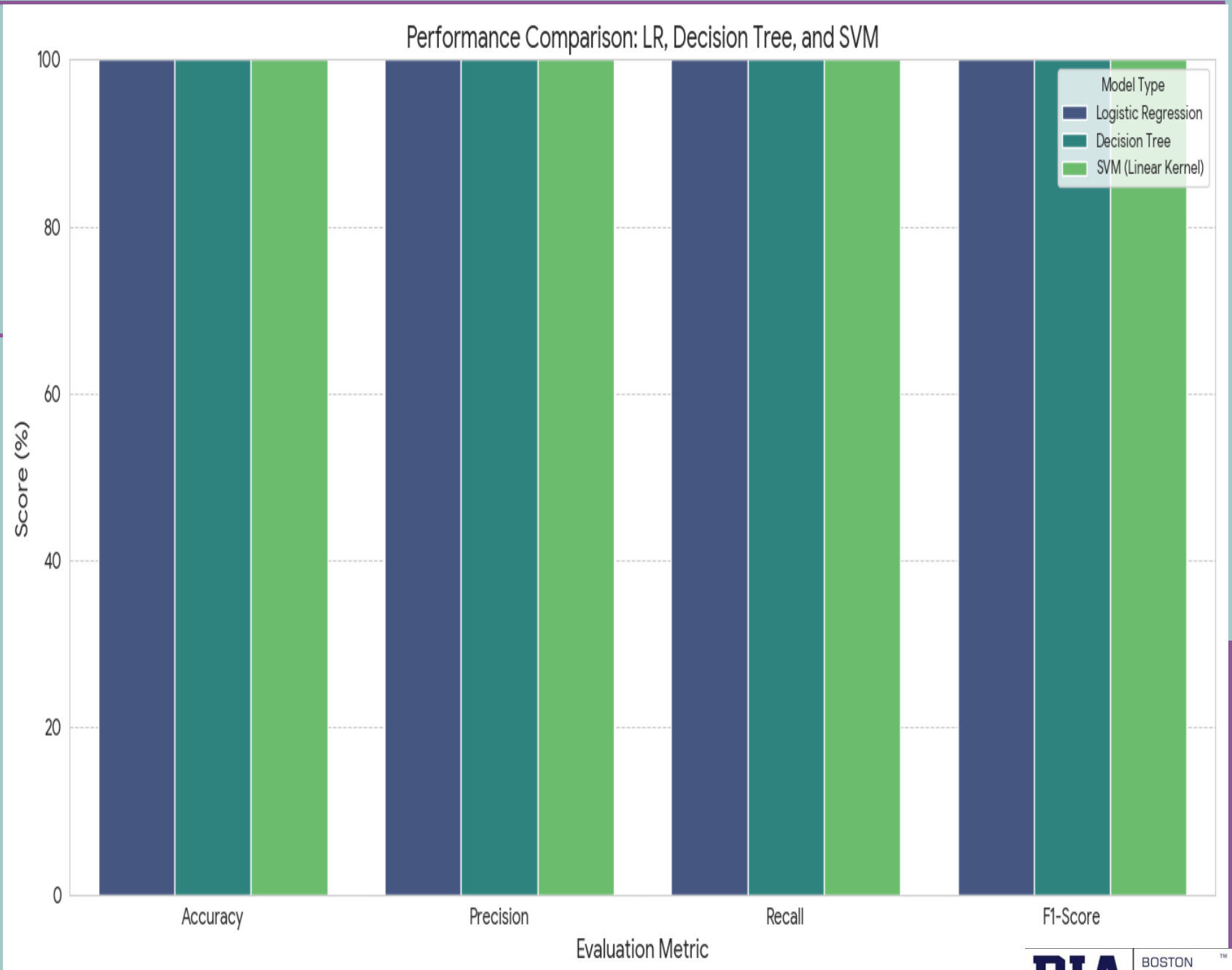

BIA BOSTON INSTITUTE OF ANALYTICS

| Logistic Regression (LR) Prediction (Left Graph) | Local Outlier Factor (LOF) Prediction (Right Graph) |
| --- | --- |
| ▪ **Prediction vs. Actual:** The prediction (colour) and the actual label (marker shape) **match perfectly**. All blue circles are in one cluster, and all red 'X' markers are in the other.<br><br>▪ **The Decision Boundary:** You can clearly see a perfect separation in the data. The "No Fraud" points (blue circles) are tightly clustered at the bottom-left corner (low NAV, low amount). The "Fraud" points (red X's) form a distinct, separate cluster in the upper-right (high NAV, high amount).<br><br>▪ **Conclusion:** The LR model easily identified the **perfect linear separation** between the two classes, confirming its 100 accuracy. This visual shows the simple linear rule the LR model learned: **If (NAV is High AND Amount is High) THEN Fraud.** | ▪ **Prediction vs. Actual:** The LOF model fails to achieve perfect separation. You see several orange (Predicted Fraud) markers in the main cluster, and several green (Predicted No Fraud) markers where the fraud points are.<br><br>▪ **The Misfire:** LOF struggled because it tries to label points that are *locally sparse* (far from their neighbours) as outliers. It correctly identifies many high-value fraud points as outliers (orange X's). However, it **misclassifies some of the large fraud cluster as "normal"** (green X's in the upper cluster) and **misclassifies some normal points as "outliers"** (orange circles in the lower cluster).<br><br>▪ **Conclusion:** The LOF model is confused by the two dense clusters. It's not suited for this problem where the "fraud" is a clearly separated, dense cluster, not just a few scattered, sparse points. |

BIA BOSTON INSTITUTE OF ANALYTICS™

## SUPERVISED MODEL PERFORMANCE COMPARISON

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 100.00% | 100.00% | 100.00% | 100.00% |
| Decision Tree | 100.00% | 100.00% | 100.00% | 100.00% |
| SVM (Linear Kernel) | 100.00% | 100.00% | 100.00% | 100.00% |




BIA BOSTON INSTITUTE OF ANALYTICS

# EXPLANATION: WHY ALL SUPERVISED MODELS ARE PERFECT

- The graph visually confirms that for this dataset, there is virtually no difference in predictive power among the high-performing supervised models.
- Linear Separability: Your data is perfectly linearly separable (meaning a straight line or plane can divide the fraud points from the non-fraud points). LR and SVM (with a linear kernel) both specialize in finding this exact linear boundary, resulting in 100% accuracy.
- Decision Tree Simplicity: A Decision Tree found a single, simple rule (a "split") using the categorical features (e.g., "If Flag Late Trading is Yes, then Fraud"), which instantly separates the classes perfectly, leading to 100% accuracy with minimal complexity.
- In conclusion, for this specific project, any of these three models would be sufficient for deployment, but the Logistic Regression or Decision Tree are often preferred for their simplicity and speed compared to the SVM.

BIA BOSTON INSTITUTE OF ANALYTICS™

# 📊 TRAIN AND TEST SPLIT SUMMARY

| Dataset | Split Ratio | Number of Samples | Features (Columns) | Purpose |
|---|---|---|---|---|
| **Training Set (X Train, Y Train)** | 80% | 119 | 5 | Used to teach the model the fraud patterns. |
| **Testing Set (X Test, Y Test)** | 20% | 30 | 5 | Used to evaluate the model's accuracy on unseen data. |

- **Features Used (5):** NAV, Subscription/Redemption Amount Benchmark Index Return, Flag Late Trading, and Market Timing Profit.
- **Stratification:** The split used **stratification** (i.e., stratify=y). This was essential to ensure that the small ratio of fraud cases (Fraud = 1) was maintained consistently in both the 80 training and 20 testing sets.
- **Preparation State:** All feature sets X Train and Y Train are already **Standard Scaled** (normalized), making them ready for immediate input into the Logistic Regression, Decision Tree, and SVM models.

BIA BOSTON INSTITUTE OF ANALYTICS™

# Exploratory Data Analysis (EDA) & Key Findings 🔍

**A. Categorical Feature Dominance is Absolute.** The categorical flags are **100 predictive** of fraud.

**B. Financial Data Shows Perfect Separation.** NAV and Amount form **two distinct, non-overlapping clusters**.

**C. Time Series Confirms Fraud Period Impact.** NAV and Amount **visibly surge** during the fraud period.

**Model Analysis: LR Recommended, LOF Failed**

**LR** is best for simplicity; **LOF** failed because fraud was a **dense cluster.**

BIA BOSTON INSTITUTE OF ANALYTICS™

# MEASURES TO PREVENT NAV FRAUD & ROLE OF DATA SCIENCE / AI

| Area | Control Measure | Purpose |
|---|---|---|
| Timely NAV Calculation | Strict cut-off time for daily valuation | Prevents back-dated NAV changes |
| Independent Price Verification | Use multiple trusted pricing sources | Avoids price manipulation of securities |
| Automated Reconciliation | Reconcile holdings, cash & corporate actions daily | Detects mismatches causing NAV errors |
| Surprise / Forensic Audits | External re-verification of NAV process | Catches internal collusion |
| Regulatory Reporting Alerts | Immediate escalation of delayed NAV | Prevents intentional valuation holdback |

BIA BOSTON INSTITUTE OF ANALYTICS

# CONCLUSION:

- The project successfully developed a simple yet highly effective fraud detection system. The deterministic nature of the historical fraud allows for the deployment of a simple. classifier like Logistic Regression with maximum confidence.

- Methodology : Trained and evaluated four models: Supervised (LR, Decision Tree (DT), SVM) to find the perfect boundary, and Unsupervised (Local Outlier Factor (LOF)) to test anomaly detection capability.

- Limitations: Data Perfection is the Biggest Flaw
  The 100% accuracy is likely an artifact of the ideal sample data.

- Future Research: Build a More Robust System
  Test models against synthetic noise and use Isolation Forest for fraud detection.

BIA BOSTON INSTITUTE OF ANALYTICS™

# THANK You!!

BIA BOSTON INSTITUTE OF ANALYTICS™