Data Understanding for Synthetic Fraud Generation: A Diffusion Model Approach to E-commerce Fraud

Akash Murali, Anish Rao, Raghu Ram Sattanapalle

January 28, 2025

Source

The dataset used for this analysis is the *'Fraud Ecommerce' dataset*, obtained from Kaggle.

- **Link:** `https://www.kaggle.com/datasets/vbinh002/fraud-ecommerce`
- **Access:** Publicly available with no restrictions.

Structure and Metadata

Key features identified in the dataset include:

- **purchase_value:** Value of the purchase.
- **source:** User source (e.g., SEO, Direct, or Ads).
- **browser:** Browser used by the user.
- **time_diff:** Time difference between signup and purchase.
- **age:** Age of the user.

Dataset Size: The dataset contains 151,112 rows and 12 columns.

Table 1: Statistical Summary of the Dataset

Feature	Count	Mean	Std	Min	25%	Max
user_id	151112	200171.04	115369.29	2.00	100642.50	400000.00
purchase_value	151112	36.94	18.32	9.00	22.00	154.00
age	151112	33.14	8.62	18.00	27.00	76.00
ip_address	151112	2.15×10^9	1.25×10^9	5.21×10^4	1.09×10^9	4.29×10^9
class	151112	0.09	0.29	0.00	0.00	1.00
time_diff	151112	1370.01	868.41	0.00	607.43	2879.99

Missing Data

- No missing values were found in the dataset.

Anomalies

- **Time Difference:** Fraudulent transactions tend to occur shortly after signup, whereas non-fraudulent transactions, on average, occur later.

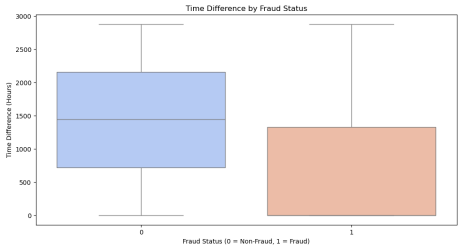

Figure 1: Time difference between signup and purchase. Fraudulent transactions are skewed toward shorter time intervals.

- The dataset is logically consistent, with no negative transaction values or future dates.

Bias

- **Class Imbalance:** The dataset exhibits class imbalance, with only around 9.4% of transactions being fraudulent.

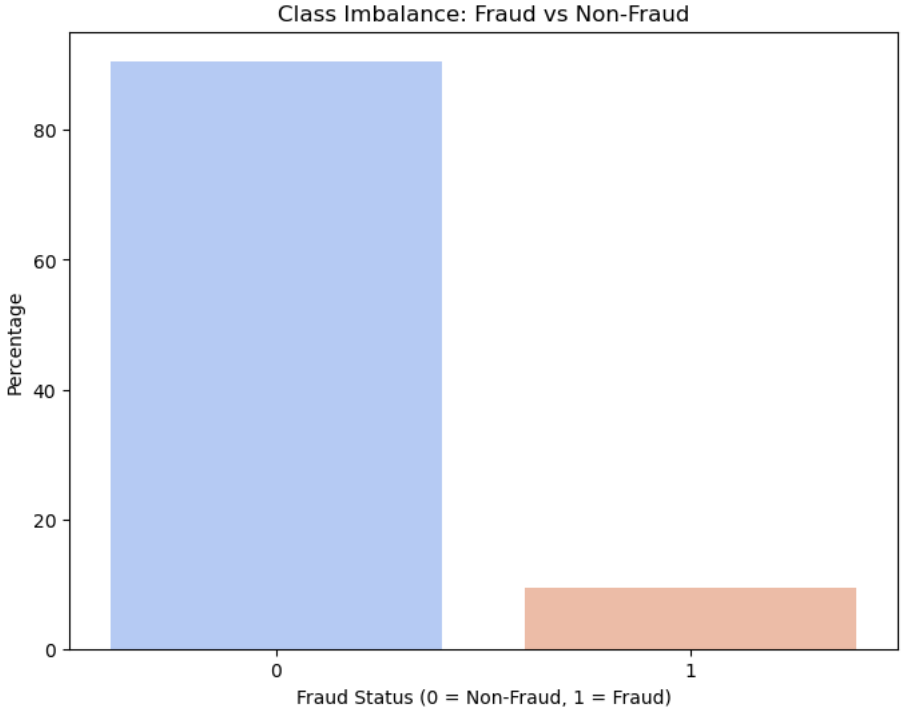

Figure 2: Class imbalance: Non-fraudulent transactions significantly outnumber fraudulent ones.

- **Source Bias:** Fraudulent activity varies significantly across user acquisition sources, with some sources showing higher fraud rates.

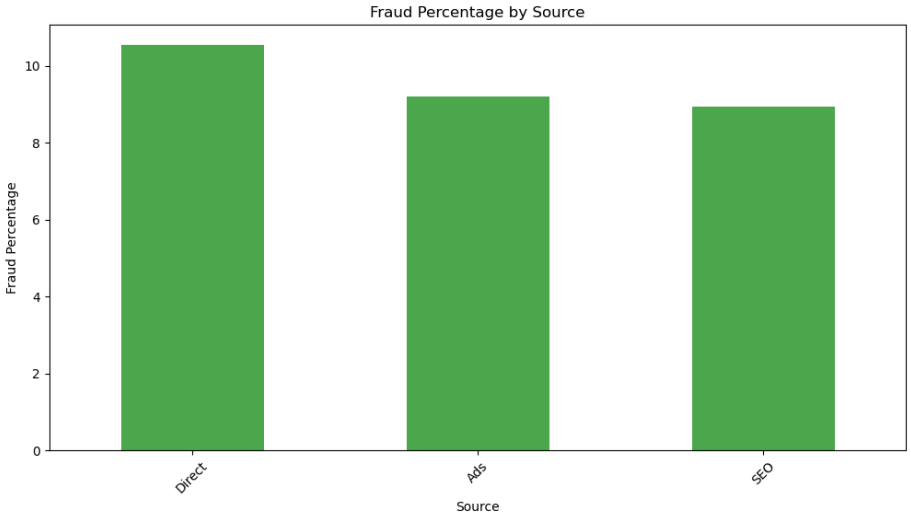

Figure 3: Fraud percentage by user source. Certain sources show higher fraud rates.

- **Demographic Bias:** Both genders are represented, but fraudulent activity seems slightly higher in one group.
- **Age Bias:** Younger users (ages 18–45) dominate the dataset. Fraudulent and non-fraudulent transactions have similar peaks in the age distribution.

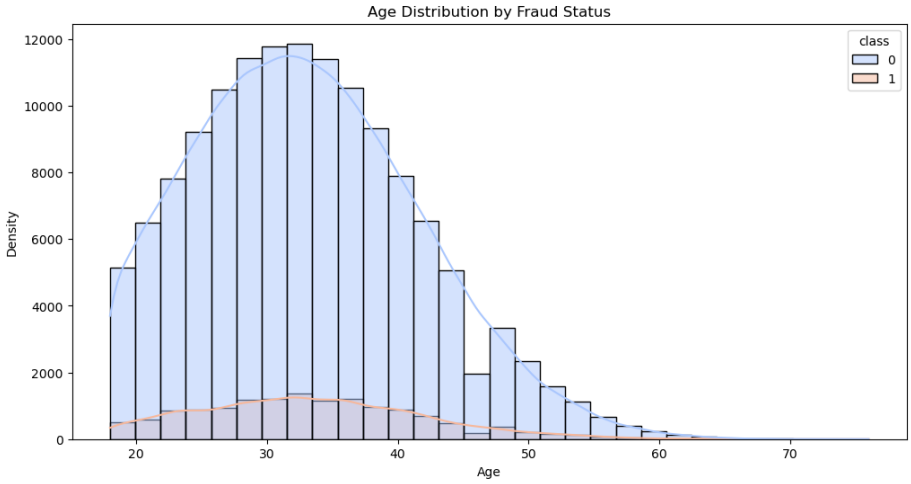

Figure 4: Age distribution of users. Fraudulent and non-fraudulent transactions share similar peaks.

- **Browser Bias:** Fraudulent activity is slightly higher in Chrome and Firefox compared to other browsers.

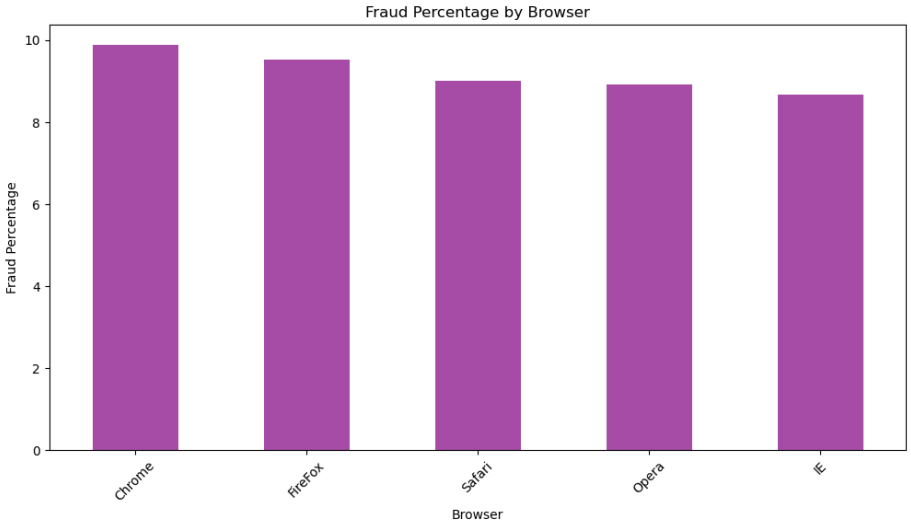

Figure 5: Fraudulent activity by browser. Chrome and Firefox exhibit higher fraud rates.

Distributions

- **Purchase Value:** For both fraud and non-fraud, the distribution is right-skewed, with most transactions at lower values and a few high-value outliers.

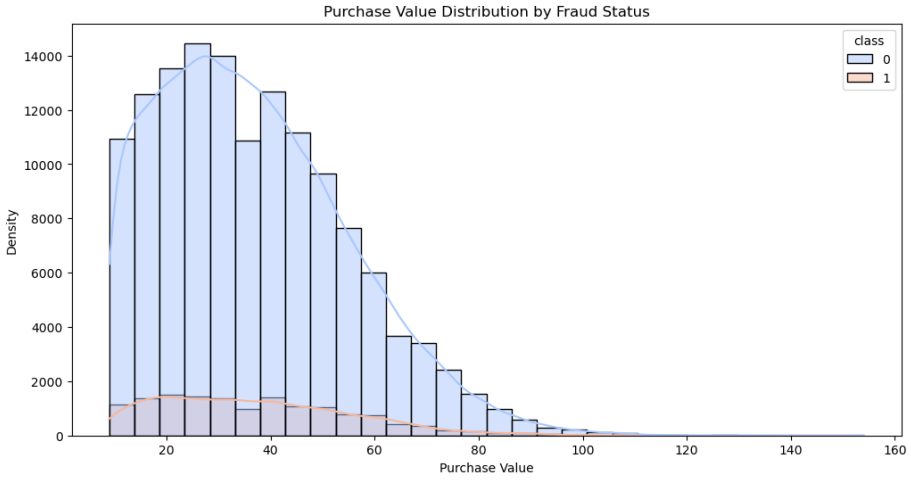

Figure 6: Purchase value distribution. Most transactions are concentrated at lower values.

- **Age:** Mostly normally distributed with a slight right skew.
- **Time Difference:** Fraudulent transactions are skewed towards shorter time differences between signup and purchase. This can be seen in figure 1.

Feature Correlation

- Weak correlations between most features, except for a moderate negative correlation between *time_diff* and the fraud class.

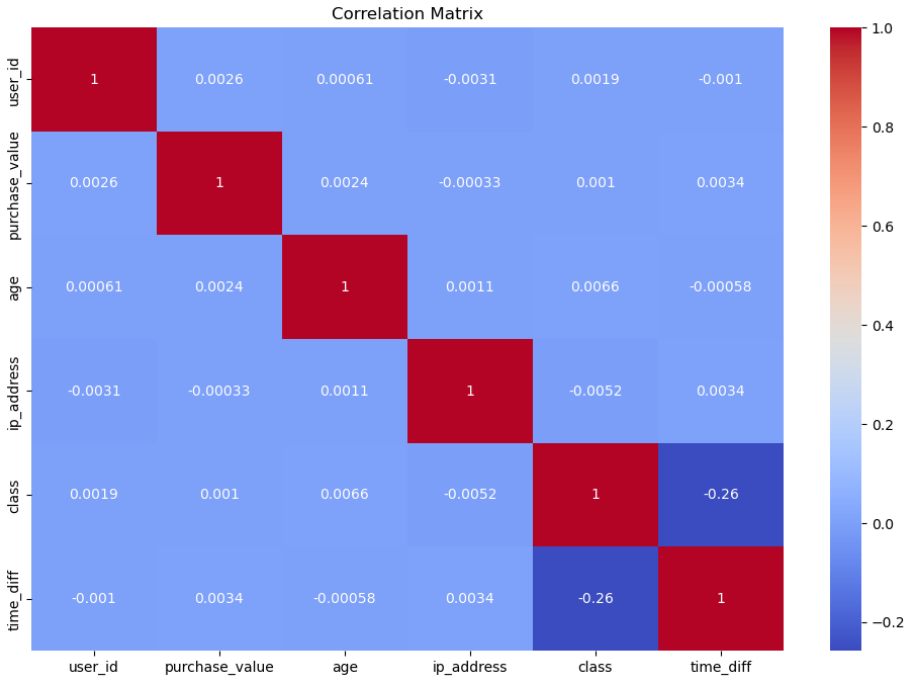

Figure 7: Correlation matrix. A moderate negative correlation exists between *time_diff* and fraud class.

Categorical Data

Unique Categories in Categorical Variables

The dataset contains the following unique categories for each categorical variable:

Variable	Unique Categories
Source	3
Browser	5
Sex	2

Table 2: Number of unique categories in each categorical variable.

Category Distribution

Category Distribution for `source`

The distribution of the `source` feature is as follows:

Source	Percentage
SEO	40.11%
Ads	39.63%
Direct	20.26%

Table 3: Distribution of `source` feature.

Category Distribution for `browser`

The distribution of the `browser` feature is as follows:

Browser	Percentage
Chrome	40.65%
IE	24.30%
Safari	16.32%
FireFox	16.29%
Opera	2.43%

Table 4: Distribution of `browser` feature.

Category Distribution for `sex`

The distribution of the `sex` feature is as follows:

Sex	Percentage
Male (M)	58.43%
Female (F)	41.57%

Table 5: Distribution of `sex` feature.

Category Distribution by Fraud Class

Source

Source	Fraud Transactions (%)	Non-Fraud Transactions (%)
Ads	38.96	39.70
SEO	38.24	40.31
Direct	22.80	20.00

Table 6: Category distribution for `source` by fraud class.

Browser

Browser	Fraud Transactions (%)	Non-Fraud Transactions (%)
Chrome	42.89	40.42
IE	22.52	24.49
FireFox	16.55	16.26
Safari	15.72	16.39
Opera	2.32	2.44

Table 7: Category distribution for `browser` by fraud class.

Sex

Sex	Fraud Transactions (%)	Non-Fraud Transactions (%)
Male (M)	59.60	58.31
Female (F)	40.40	41.69

Table 8: Category distribution for `sex` by fraud class.

- The `sex` and `browser` features are imbalanced, with certain categories dominating the distribution.
- The `source` feature shows a relatively balanced distribution across fraud and non-fraud transactions.
- Fraudulent transactions show slightly different category preferences compared to non-fraudulent transactions, which can be leveraged for fraud detection.

Ethical Concerns

- **Reidentification Risks:** The combination of user identifiers (e.g., user_id, ip_address, device_id) increases the risk of reidentifying individuals, especially when used alongside external datasets.
- **Bias and Discrimination:** Using demographic features like age and sex in fraud detection models could lead to biased outcomes, disproportionately impacting specific groups.
- **Security Vulnerabilities:** Publishing or exposing ip_address or device_id data could make users vulnerable to cyber threats such as targeted hacking or tracking.

Alignment with Goals

The dataset aligns well with the project's primary objective of detecting fraudulent transactions and addressing class imbalances. It provides a comprehensive set of attributes that are crucial for fraud analysis, including:

- **Transaction Details:** Attributes such as transaction amount (`purchase value`), timestamp (`purchase time`), and signup time details.
- **User Attributes:** Information about the user, including `age`, `sex`, and `user_id`, which help in understanding user profiles.
- **Device and Network Attributes:** Attributes such as `device_id` and `ip_address` provide insights into transaction patterns and potential anomalies.

The dataset contains several features that are essential for effective analysis and modeling:

- **Temporal Features:** The `purchase time` attribute enables the extraction of temporal patterns such as transaction hour and day of the week, which are critical for detecting fraudulent behaviors.
- **Categorical Features:** Features like `source`, `browser` provide categorical information that can be encoded and used in machine learning models.
- **Demographic Features:** Attributes such as `age` and `sex` offer demographic insights that can aid in understanding fraud trends.
- **Geographic Features:** The inclusion of `ip_address` enables geographic analysis of transactions, identifying location-based fraud patterns.

Scalability

The dataset consists of 151,112 rows and 11 columns, making it manageable on most modern computing systems, including laptops and desktops with at least 4-8 GB of RAM. Given its size, standard data processing tools such as Pandas and NumPy can efficiently handle it without significant memory constraints.

Model Training Considerations

For training machine learning models:

- The dataset can be processed and trained locally on machines with moderate hardware specifications.
- Cloud-based services such as Google Colab, AWS, or Azure can be utilized for more intensive computations or scalability purposes.
- Feature engineering and preprocessing can be performed efficiently using batch processing methods without requiring distributed computing frameworks like Apache Spark.

Transformations

Numerical features in the dataset require preprocessing to improve model performance:

- **Normalization:** Ensures that numerical features are scaled between a fixed range, making them suitable for distance-based models such as k-Nearest Neighbors (k-NN) and tree based models.
- **Standardization:** Centers numerical features by subtracting the mean and scaling to unit variance, which benefits models like Logistic Regression and Support Vector Machines (SVM).

We would apply suitable transformation technique based on the model we choose to train our data.

Feature Engineering Opportunities

Several new features can be derived to enhance fraud detection:

- **Time-based Features:** Extracts temporal patterns from timestamps, such as the hour of signup and the hour of purchase, to analyze user behavior.
- **Time Difference:** Measures the duration between signup and purchase in hours. A significantly short time interval may indicate fraudulent activity.

- **Day of the Week and Hour of Purchase:** Identifies patterns in fraudulent transactions by tracking when purchases occur. Certain times or days may have a higher fraud rate.

Data Encoding

Categorical variables require appropriate encoding techniques to be effectively used in machine learning models:

- **Encoding Choice for This Dataset:** Since all categorical features in this dataset have low cardinality, one-hot encoding is sufficient to represent them without loss of information.

Predictive Power

- **Feature Predictive Analysis**:
  - Numerical Features:
    - `time_diff`: Strongest predictor
      - Fraud cases: mean 28 days after signup
      - Non-fraud cases: mean 60 days after signup
      - Some fraud cases occur almost immediately (0.02 minutes)
    - `purchase_value`: Similar distributions
      - Fraud mean: $36.99
      - Non-fraud mean: $36.93
      - Fraud transactions capped at $111 (vs $154 for non-fraud)
    - `age`: Limited predictive power
      - Fraud mean: 33.32 years
      - Non-fraud mean: 33.12 years
      - Fraud age range more concentrated (18-68 vs 18-76)
  - Device and IP Patterns:
    - 1,044 devices used in multiple frauds
    - 759 IPs used in multiple frauds
    - Significant reuse patterns indicating organized fraud
  - Temporal Patterns:
    - Higher fraud rates during early morning hours
    - Clear hourly pattern in fraud occurrence
    - Time-based features show strong predictive power
- **Feature Selection Requirements**:
  - Primary Predictors:

    - Time-based features (signup to purchase gap)
    - Device and IP reuse patterns
    - Source-specific fraud rates
  - Dimensionality Reduction Needed:
    - High-cardinality features (`device_id`, `ip_address`)
    - Temporal features requiring transformation
    - Categorical variables needing encoding

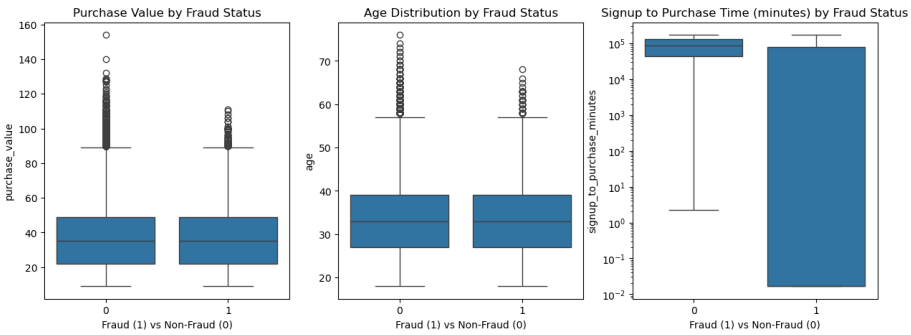

Figure 8: Distribution of Purchase Value, Age, and Signup-to-Purchase Time by Fraud Status

Target Variable

- **Definition**:
  - Binary classification problem:
    - Target variable: `class`
    - 0: Non-fraudulent transaction
    - 1: Fraudulent transaction
  - Well-defined characteristics for diffusion modeling:
    - Clear binary distinction for conditional generation
    - No missing or ambiguous cases
    - Strong feature relationships to preserve
- **Class Distribution for Diffusion Training**:
  - Current distribution:
    - Non-fraud (0): 90.6% (136,904 transactions)
    - Fraud (1): 9.4% (14,208 transactions)

  - Advantages for diffusion modeling:
    * 14,208 fraud cases provide sufficient examples for learning patterns
    * More balanced than typical fraud datasets (usually $< 1\%$ fraud)
    * Enough samples to learn feature correlations within fraud class

- **Generation Requirements**:
  - Pattern preservation needs:
    * Maintain temporal relationships (signup-to-purchase patterns)
    * Preserve demographic distributions within fraud class
    * Keep realistic purchase value ranges ($9.00 - $111.00 for fraud)
  - Quality metrics for generated samples:
    * Distribution similarity to real fraud cases
    * Feature correlation preservation
    * Realistic categorical variable distributions (source, browser, sex)

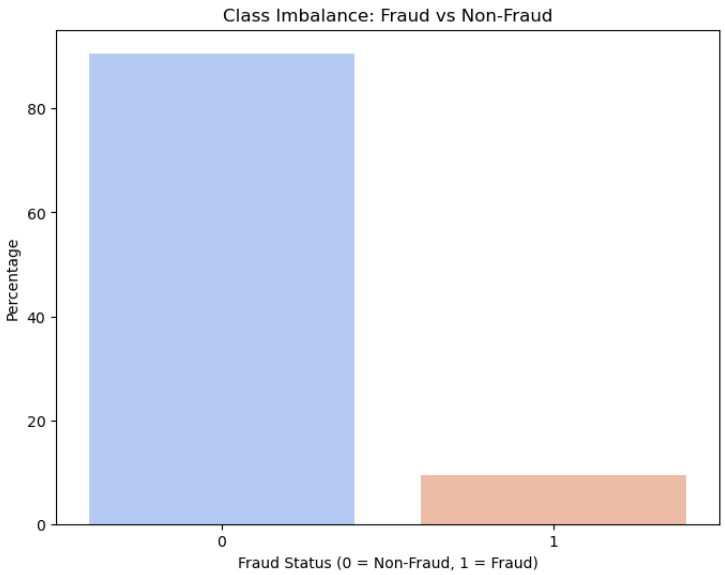

Figure 9: Distribution of fraud vs non-fraud transactions showing class imbalance

Validation Strategy

- **Dataset Splitting Strategy**:

  - For Diffusion Model Training:
    * Training: 100% of data used to learn fraud patterns
    * Rationale: Need all available fraud cases (14,208) to learn complete distribution
    * No traditional split needed as we're generating, not predicting
  - For Evaluating Generated Samples:
    * Compare synthetic samples against real data distributions
    * Use statistical tests for distribution matching
    * Assess feature relationship preservation

- **Dependencies to Preserve**:
  - Temporal Dependencies:
    * Signup-to-purchase time gaps (mean 28 days for fraud)
    * Hour-of-day fraud patterns
    * Purchase timing relationships
  - Feature Dependencies:
    * Device-IP relationships (1,044 devices, 759 IPs in multiple frauds)
    * Source-specific fraud rates
    * Demographic patterns within fraud cases

Data Leakage

- **Traditional Leakage Concerns**:
  - Not applicable in conventional sense because:
    * We're using full dataset to learn fraud patterns
    * Goal is generation, not prediction
    * No train/test split in traditional sense

- **Relevant Leakage Risks**:
  - Identity Information:
    * `user_id`: Must be excluded from generation
    * `device_id`: Should generate new, but maintain reuse patterns
    * `ip_address`: Should generate new, while preserving geographic patterns
  - Temporal Information:
    * Absolute timestamps should not be copied
    * Need to generate new, plausible timestamps
    * Preserve time gaps and patterns without exact replication

- **Prevention Strategies**:
  - For Generated Data:
    - Generate new identifiers rather than copying
    - Maintain statistical patterns without replicating exact values
    - Preserve relationships while creating novel combinations
  - For Downstream Evaluation:
    - Use different evaluation periods for synthetic data testing
    - Ensure generated samples don't replicate exact real cases
    - Validate on pattern preservation, not exact matching

Interpretability

- **Stakeholder Insights from Data**:
  - Fraud Patterns:
    - Time-based insights: Fraud occurs 28 days after signup vs 60 days for legitimate
    - Channel-specific risks:
      - Direct: Highest risk (10.54%)
      - Ads: Medium risk (9.21%)
      - SEO: Lowest risk (8.93%)
    - Value patterns: Fraud transactions capped at $111 vs $154 for legitimate
- **Communication Methods**:
  - Quality of Generated Data:
    - Distribution comparisons (real vs. synthetic):
      - Purchase value distributions
      - Time gap patterns
      - Demographic distributions
    - Visual validations:
      - Box plots comparing real vs. generated fraud patterns
      - Time series plots of fraud rates
      - Feature correlation heatmaps
  - Business Metrics:
    - Pattern preservation rates
    - Feature relationship maintenance
    - Fraud detection improvement metrics
- **Key Visualizations**:

  - Pattern Validation:
    * Purchase value distributions (real vs. generated)
    * Temporal pattern preservation
    * Source-specific fraud rate matching
  - Quality Assessment:
    * Feature correlation preservation
    * Category distribution matching
    * Time-gap pattern reproduction

Limitations

- **Dataset Limitations for Diffusion Modeling**:
  - Sample Size Constraints:
    * Only 14,208 fraud cases to learn patterns from
    * Limited examples of diverse fraud behaviors
    * May affect quality of generated patterns
  - Feature Limitations:
    * Missing modern fraud indicators:
      · No device fingerprinting
      · Limited IP information
      · No behavioral patterns over time
    * Basic transaction details:
      · Single purchase value only
      · No transaction categories
      · No merchant information
  - Temporal Coverage:
    * Single year of data (2015)
    * May miss seasonal fraud patterns
    * Outdated compared to current fraud techniques
- **Desired Additional Data**:
  - Enhanced Transaction Features:
    * Detailed product information
    * Transaction categories
    * Merchant risk scores
    * Payment method details
  - Technical Indicators:
    * Device fingerprinting data
    * Network/IP geolocation

    - Browser fingerprinting
    - Session behavior patterns
- Historical Context:
    - User transaction history
    - Account age and activity
    - Previous fraud attempts
    - Device/IP history
- Modern Patterns:
    - Recent fraud techniques
    - Multiple years of data
    - Seasonal pattern information