A CASE STUDY ON FRAUD DETECTION

Abhisek Kumar Gupta

MQMS2301
ISI Bangalore

DETECTING FRAUDULENT TRANSACTIONS USING UNSUPERVISED LEARNING

❖ Context:

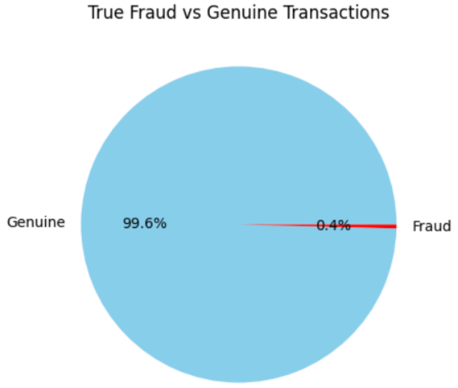
Identify anomalous transactions that could represent fraudulent behavior using unsupervised models.

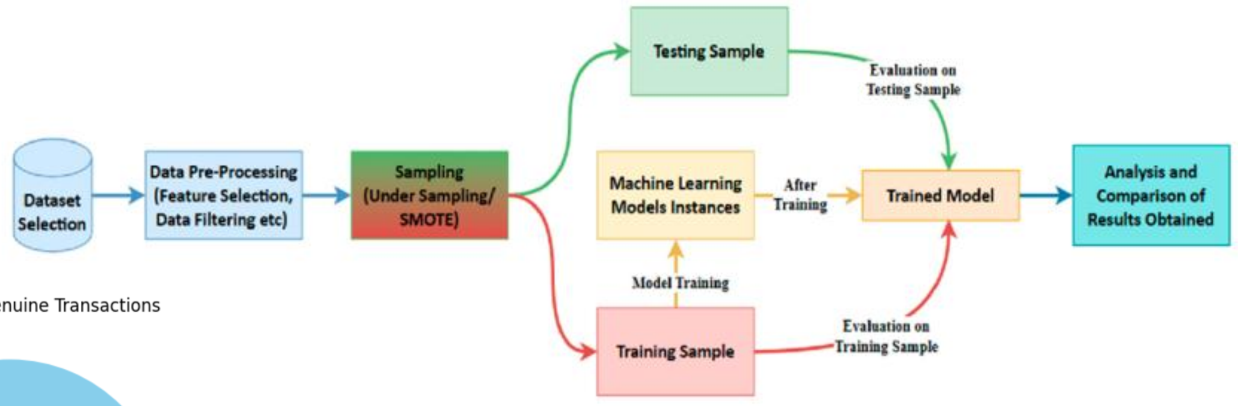
❖ Objectives:

- Flag unusual transactions.
- Understand model decision-making (via SHAP).
- Reduce false positives and false negatives.
- Compare models based on suitable metric.

❖ Datasets:

- Train Data:
  - (1296675, 22), no missing values & duplicates
- Test Data:
  - (555719, 22), no missing values & duplicates





Flow Diagram of Credit Card Fraud Detection using Machine Learning.

Source: https://www.researchgate.net/figure/Flow-Diagram-of-Credit-Card-Fraud-Detection-using-Machine-Learning_fig1_376523310

MODELS USED FOR ANOMALY DETECTION

Choice of the models:

1. **Isolation Forest** – random partitioning of feature space.
2. **Local Outlier Factor** – density-based detection.
3. **Autoencoder** – neural net that flags high reconstruction error.

Preprocessing:

Feature Engineering: Key Steps Implemented

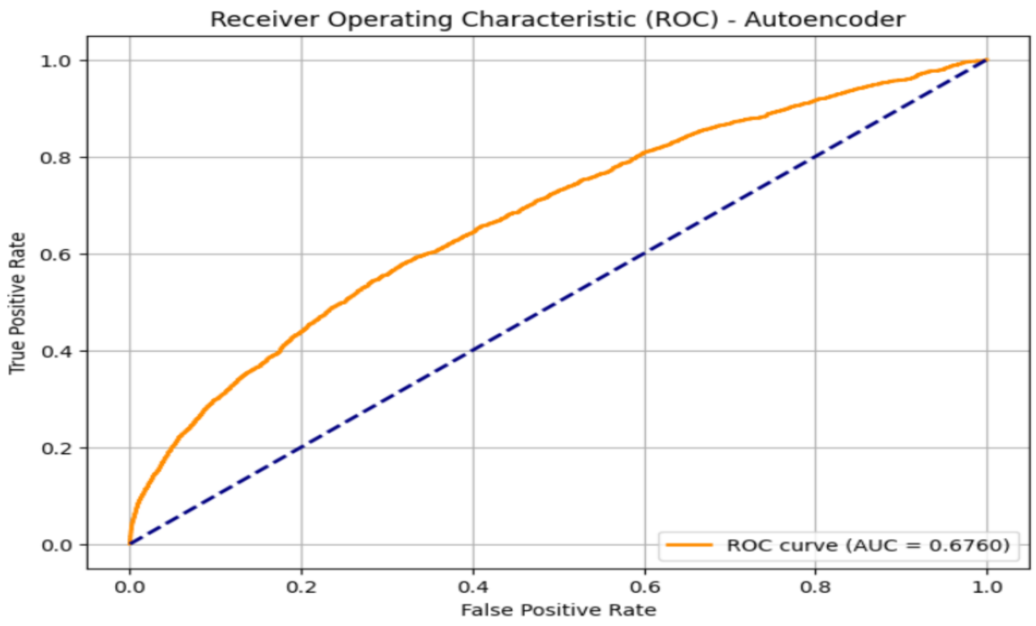
1. **Derived Features**
   a. Age: Calculated from dob (date of birth) and transaction timestamp.
   b. Temporal Features: hour, day, month, weekday extracted from trans_date_trans_time.
2. **Data Cleaning:** Dropped Irrelevant Columns:
   a. High-cardinality identifiers: trans_num, first, last, street.
   b. Redundant timestamps: dob, trans_date_trans_time.
3. **Categorical Encoding**
   a. Label Encoding: Applied to all categorical columns (e.g., merchant, category, gender).
4. **Final Dataset**
   a. Train/Test Sets: Processed datasets with engineered features, ready for anomaly detection.
   b. Data scaled using StandardScaler.
5. **Predictions mapped to binary:** 1 → fraud, 0 → normal.

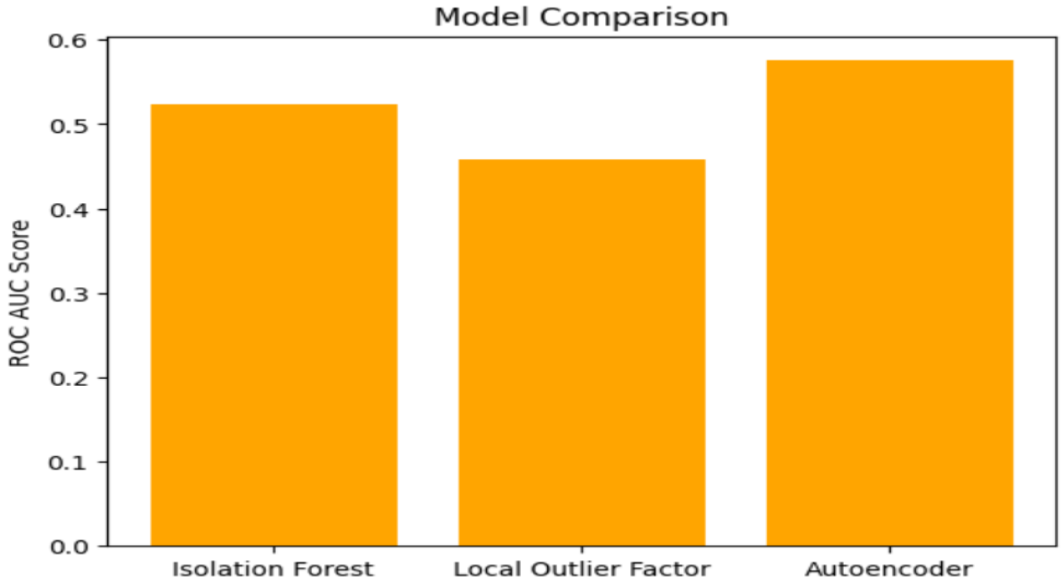
Aspect	Isolation Forest (IF)	Local Outlier Factor (LOF)	Autoencoder (AE)
Algorithm Type	Tree-based ensemble	Distance-based, unsupervised	Neural network (deep learning-based)
Key Parameters	n_estimators, contamination, max_samples	n_neighbors, contamination	hidden layer sizes, activation, epochs, batch_size
Detection Logic	Isolates observations in trees; anomalous points get isolated quickly → fewer splits → lower average path length = more anomalous	Compares local density of a point to its neighbors. Lower local density = more likely to be an outlier	Learns to reconstruct normal patterns; high reconstruction error = anomaly
Anomaly Score	Based on average path length in isolation trees	Based on the ratio of local densities	Based on reconstruction error between input and output
Strengths	Fast, handles high dimensions well	Effective for local outliers, intuitive	Captures complex nonlinear relationships
Weaknesses	May miss local anomalies	Sensitive to n_neighbors, doesn't scale well to large datasets	Requires more data, training time, and tuning
Explainability	SHAP works reasonably well with trees	Partial SHAP support via surrogate model	SHAP support via surrogate model (e.g., decision tree)
Scaling Required	Not always, but recommended	Yes	Yes (essential for neural networks)
Typical Use Cases	General-purpose anomaly detection (e.g., fraud, intrusion)	Detecting local anomalies in dense clusters (e.g., sensor faults)	Complex fraud patterns, time-series anomaly detection

MODEL EVALUATION

Choice of Metric: ROC-AUC

- **Class Imbalance Handling:** Anomalies are usually rare (e.g., less than 1% of the data).
- Metrics like accuracy can be misleading (e.g., 99% accuracy by predicting all normal).
- ROC-AUC evaluates the model across all thresholds, **balancing:**
  - **True Positive Rate (Recall/Sensitivity)** — how well anomalies are detected.
  - **False Positive Rate** — how often normal points are misclassified as anomalies.
- It doesn't get skewed by class imbalance.





FALSE POSITIVES, FALSE NEGATIVES & BUSINESS IMPACT

False Positives (FPs) Typical causes:

- Rare but legit users flagged due to unusual but valid activity. (e.g., user travels abroad, spends more than usual).
- Customers with non-standard patterns (e.g., night-shift workers).
- Small clusters of new behavior not present in training data.
- Model behavior insight:
  - Models like Isolation Forest and LOF may overreact to legitimate rare behavior.
  - Autoencoders may flag novel but valid transactions as anomalies due to poor generalization on unseen-but-valid patterns.
- What this reveals:
  - Models are sensitive to novelty, but not all novelty is fraud.
  - Indicates need for better representation of normal behavior in training.

False Negatives (FNs) Typical causes:

- Fraud that mimics normal transaction patterns.
- Small, clever manipulations (e.g., same location, slightly different amount).
- Lack of labeled fraud examples for training.
- Model behavior insight:
  - Anomaly detection models struggle when fraud is similar to normal behavior.
  - Autoencoders may reconstruct close enough to normal → low reconstruction error.
  - IF and LOF may not find these cases as "dense" regions include some fraud-like patterns.
- What this reveals:
  - Models are blind to subtle fraud.
  - Indicates need for feature engineering or semi-supervised methods.

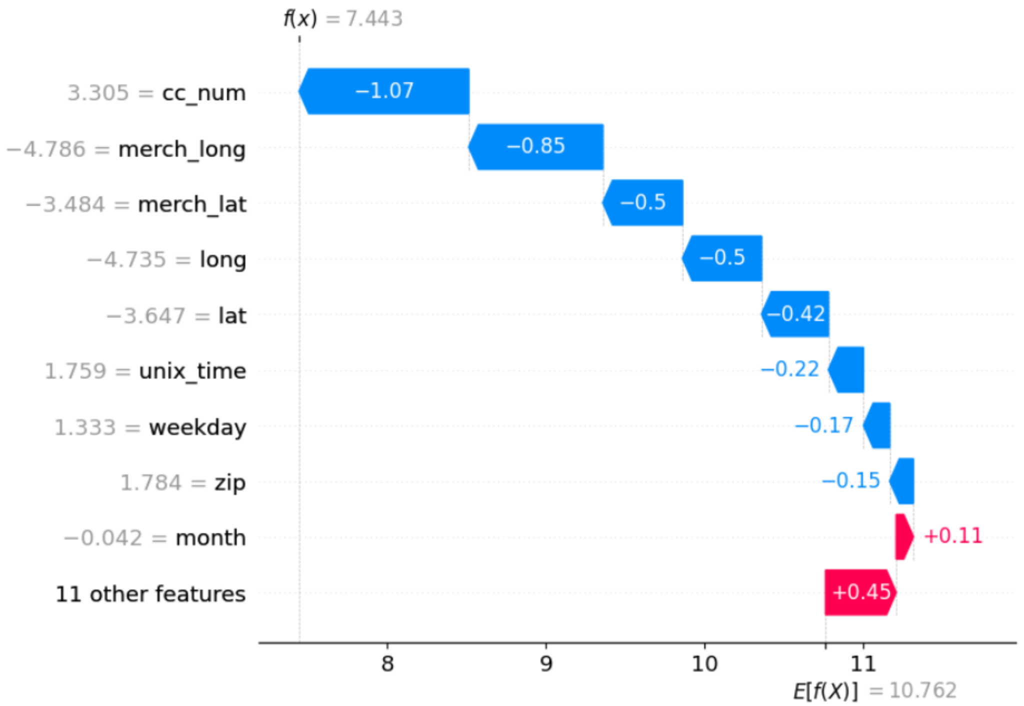
Model	Isolation Forest	Local Outlier Factor	Autoencoder
False Positives	3694 (0.66%)	316768 (57.00%)	27357 (4.92%)
False Negatives	2033 (0.37%)	1099 (0.20%)	1716 (0.31%)

EXPLAINING ANOMALIES WITH SHAP

SHAP is used on explaining the first 5 fraud detections:

For example, the first fraud detection using Isolation Forest i.e. index 864:

- The prediction value dropped from 10.762 to 7.443 due to several strong negative SHAP contributions, indicating anomaly.
- Major contributors toward fraud flagging (blue bars):
  - cc_num = 3.305: SHAP value of −1.07
  - merch_long = −4.786: SHAP value of −0.85
  - merch_lat = −3.484: SHAP value of −0.5
  - long = −4.735 and lat = −3.647: Together contributed another ~−0.92
- These suggest that this transaction occurred at unusual merchant geolocations, or that the credit card number pattern was unusual based on what the model saw in normal data.
- What made it look normal (pink bars)?
  - month = −0.042 and zip = 1.784 had positive SHAP values of +0.45 and +0.11.
- It means these made the transaction look a bit more normal or less fraudulent, but their influence was small compared to the negative contributors.



THANK YOU