Deliverable 3 Report

Notebook Code Summary
The provided code compares two anomaly detection techniques, Z-score and Isolation Forest, on a highly imbalanced dataset of credit card transactions. Here's the breakdown:

1. **Dataset Preparation**:
    - Read the dataset and performed initial exploratory data analysis (EDA) using .head(), .describe(), and .info().
    - Standardized features (excluding Time and Class) using StandardScaler for consistent feature scaling.
    - Split the data into train and test sets using train_test_split, ensuring stratification for class imbalance.
2. **Techniques Applied**:
    - **Z-score Method**:
        - Identifies outliers based on deviations from the mean.
        - Thresholds were set at 10% significance.
    - **Isolation Forest**:
        - An unsupervised learning method that isolates anomalies by randomly partitioning the feature space.
        - The model was trained on the training data and applied to the test set.
        - Negative predictions were mapped to 0 for consistency with the dataset.
3. **Metrics**:
    - Accuracy scores for both methods were calculated.
    - Confusion matrices visualized the performance of each technique.
    - Fraud detection percentage (class imbalance) was calculated as a baseline.

Findings and Comparative Analysis

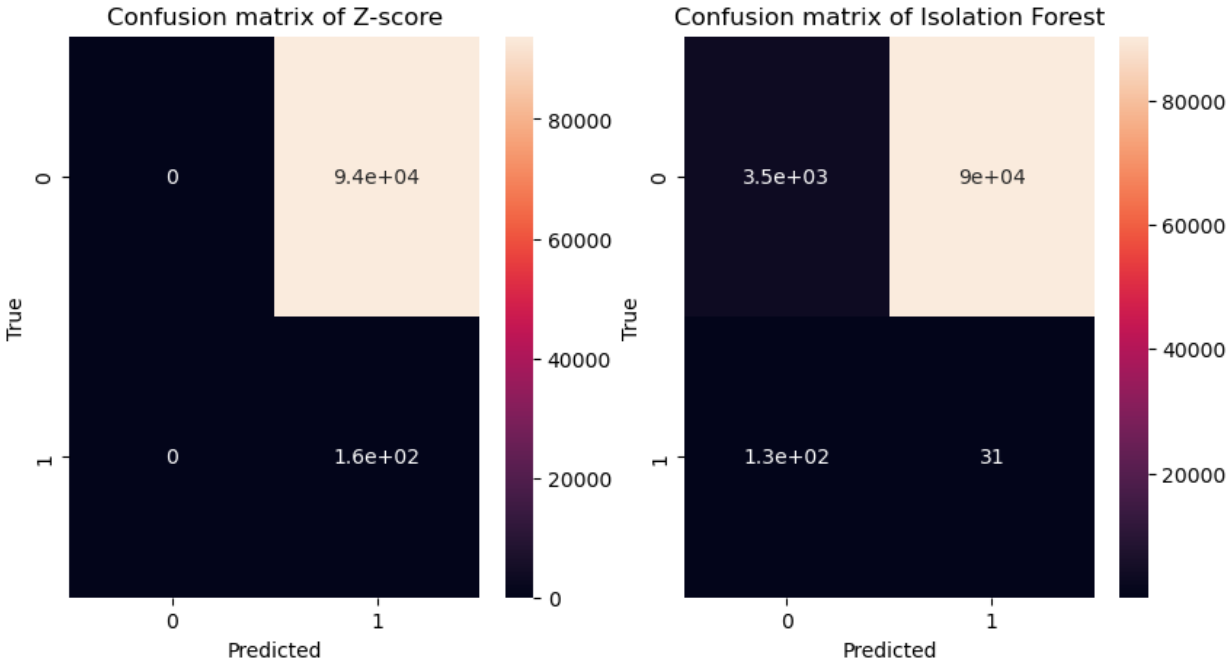
1. **Performance Metrics**:
    - **Z-score Accuracy**: 0.0017
    - **Isolation Forest Accuracy**: 0.0377
    - Fraud Percentage in Test Set: 0.0017 (highly imbalanced dataset).
2. **Confusion Matrix Insights**:

- Both methods struggled with detecting the rare fraudulent transactions.
- Isolation Forest slightly outperformed the Z-score approach but still showed poor results.

3. **Strengths and Weaknesses**:
   - **Z-score Method**:
     - Simple to implement.
     - Does not rely on training, making it computationally inexpensive.
     - Assumes Gaussian distribution, which may not suit real-world data distributions.
   - **Isolation Forest**:
     - Tailored for anomaly detection.
     - Does not assume data distribution.
     - Relies on the assumption that anomalies are sparse and differ from the majority.
4. **Scenario Fit**:
   - **Z-score**: Suitable for small datasets with normally distributed features where computational simplicity is key.
   - **Isolation Forest**: Preferred for larger, non-Gaussian datasets, especially when computational resources are available.

Insights Derived

1. Both techniques are limited in their ability to detect rare fraud cases due to the severe class imbalance.
2. A supervised approach, using models like Random Forest or Gradient Boosting, may better handle such imbalances with class weights or resampling.
3. Further preprocessing, such as feature selection and synthetic data generation (e.g., SMOTE), could improve model performance.


Confusion matrix of Z-score
0
9.4e+04
0
1.6e+02
True
0
1
Predicted
80000
60000
40000
20000
0
Confusion matrix of Isolation Forest
3.5e+03
9e+04
1.3e+02
31
True
0
1
Predicted
80000
60000
40000
20000