Mini Project 1

Ty Hawkes, Evan Miller, Talmage Hilton

2025-09-29

Abstract

Financial fraud accounts for billions of dollars in losses each year, highlighting the need for effective detection strategies. In this study, we analyze a dataset of 250,000 anonymized financial transactions collected over a two-day period, each described by 12 numerical variables. Because the dataset is unlabeled, we employ unsupervised learning methods based on outlier detection. Specifically, we apply data transformations to reduce skewness and approximate multivariate normality, enabling the use of Mahalanobis Distance to identify anomalous transactions.

Our analysis revealed that observation 130,115 had the largest Mahalanobis Distance ($\approx$ 16,233), making it the most likely fraudulent transaction. The 250th flagged transaction had a Mahalanobis Distance of 504.76. Across the 250 identified cases, the mean Mahalanobis Distance was 1,158.05, with a median of 1,078.99, suggesting a cluster of suspicious transactions with some extreme outliers. These results are consistent with the assumption that fraud manifests as statistical anomalies within otherwise regular financial data.

Introduction

In 2024, $12.5 billion dollars were lost to fraud in the United States alone, according to the Federal Trade Commission (FTC). This staggering figure underscores the critical need for effective fraud detection mechanisms. In this analysis, we will use a dataset containing 250,000 financial transactions over a two-day period to try and identify 250 of the <500 fraudulent transactions. To identify these transactions, we will use outlier detection methods under a multivariate normal distribution assumption.

Data Description

The dataset consists of 250,000 observations and 12 variables. Due to the sensitive nature of the data, the variables are anonymized and labeled as X1 to X12. Each observation represents a financial transaction, with the variables capturing various attributes of these transactions. This dataset is *not labeled*, meaning we do not have prior knowledge of which transactions are fraudulent. This lack of labeling necessitates the use of unsupervised learning techniques for fraud detection.

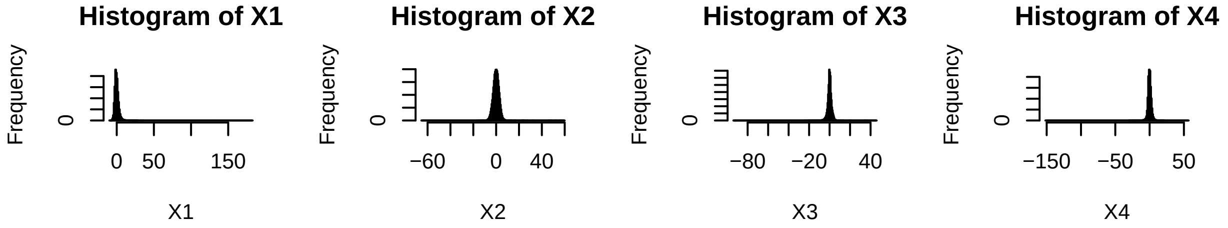

Figure 1: Example Histograms of Original Variables

Assessment of Normality

An initial assessment of the data reveals that the variables do not follow a normal distribution, as evidenced by the histograms above. In general, the distributions tend to be very heavy tailed and skewed. To improve the skewness of each variable individually, we will apply a Yeo-Johnson transformation to each variable.

The Yeo-Johnson transformation is a power transformation that can handle both positive and negative values, making it more suitable for our dataset than a Box-Cox transformation. The Yeo-Johnson transformation is defined as follows:

$$Y(\lambda) = \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda} & \text{if } y \geq 0, \lambda \neq 0 \\ \log(y+1) & \text{if } y \geq 0, \lambda = 0 \\ \frac{-(|y|+1)^{2-\lambda}-1}{2-\lambda} & \text{if } y < 0, \lambda \neq 2 \\ -\log(|y|+1) & \text{if } y < 0, \lambda = 2 \end{cases}$$

To choose a λ for an optimal transformation of each variable, the log likelihood under a normal distribution is maximized with respect to λ. To improve the normality of the distributions, we will apply the Yeo-Johnson transformation to each variable and then reassess the normality using histograms and QQ plots with confidence intervals.

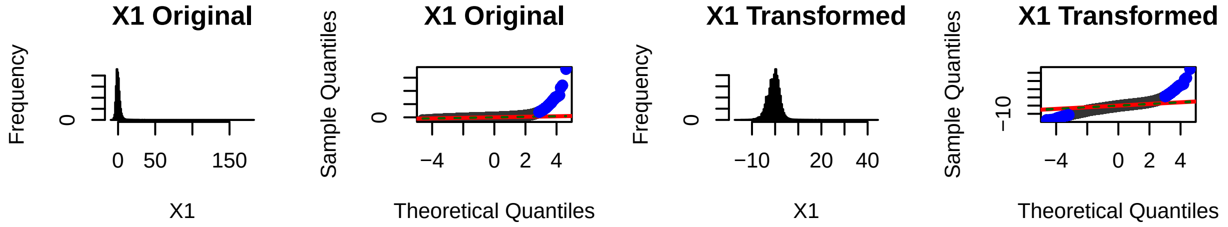

Figure 2: X1 Example of Yeo-Johnson Transformation

The figure above shows a histogram and QQ plot of X1 before and after the Yeo-Johnson transformation. Initially, X1 exhibited right skewness and very heavy tails. After applying the Yeo-Johnson transformation, the distribution of X1 appears more symmetric and less heavy-tailed. The QQ plot indicates that while the transformation has improved the fit to a normal distribution, there are still deviations from normality, particularly in the tails. While we would consider this transformation a step toward normality, it is still not perfectly normal.

Each of the other variables exhibited similar improvements in normality after the Yeo-Johnson transformation, but none achieved perfect normality. The heavy-tailed nature of the distributions persists to some extent across all variables.

To further assess the multivariate normality of the dataset, we will use Mahalanobis distance-based QQ plots. The Mahalanobis distance measures how far each observation is from the mean of the distribution, taking into account the correlations between variables. If the data were multivariate normal, the squared Mahalanobis distances would follow a Chi-square distribution with degrees of freedom equal to the number of variables. Mahalanobis distance is defined as follows:

$$D^2 = (x - \mu)^T S^{-1} (x - \mu)$$

where μ is the mean vector and S is the covariance matrix.

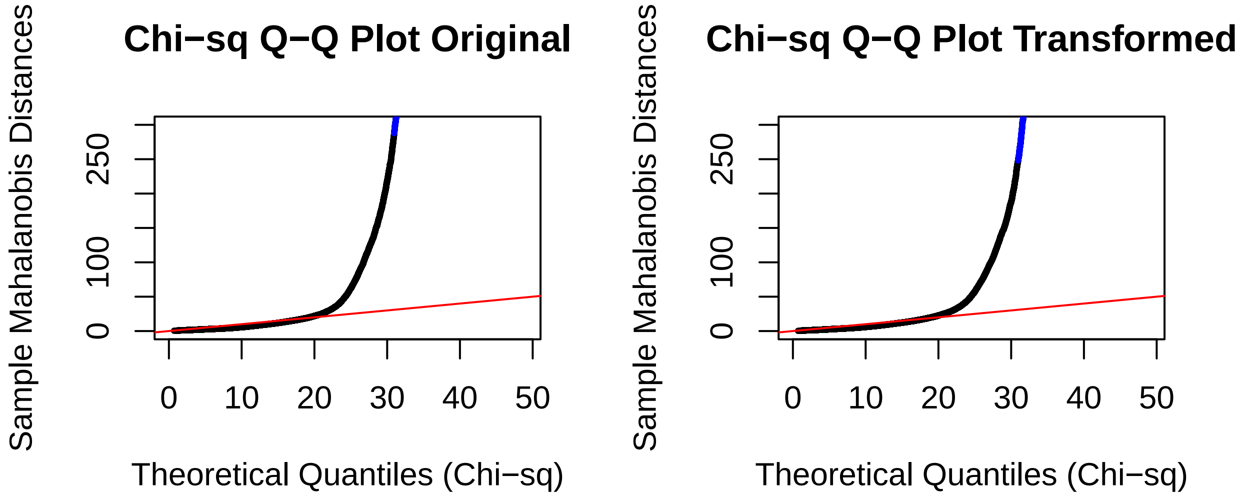

Figure 3: Original vs Yeo-Johnson Transformed QQ-Plots

Although the transformations may not appear to have improved the QQ plot, they did reduce the extremity of the 500 largest Mahalanobis distances, bringing these transactions closer to the bulk of the distribution. This suggests that the transformations mitigated some of the effects of skewness. We therefore will proceed with our analysis under the assumption of multivariate normality, acknowledging that the heavy-tailed nature of the distributions may lead to a higher rate of false positives in our outlier detection.

Outlier Detection (Potential Fraud Identification)

We identify potential fraudulent transactions in the data set by identifying outliers, assuming multivariate normality in the data. As was previously described, after transforming the data, there is still reason to believe that this data is not multivariate normal, but we will move forward with this assumption in order to employ Mahalanobis distance as a method to detect the outliers.

In our case, assuming normality, we know that Mahalanobis Distance follows a χ^2_{12} distribution. We then say that any record with a Mahalanobis Distance greater than the 99% quantile of the χ^2_{12} is an outlier. This method gave us 11739 outliers, much more than we were expecting. Again, after the transformation, there is still skewness and heavy tails, so this result is not entirely unexpected. Using the information we have, however, we believe that the transactions with the largest Mahalanobis

Distance have the greatest chance of being fraudulent. Table 1 below shows the top 10 records with the highest Mahalanobis Distance.

Table 1: Records with the highest Mahalanobis Distance

Row	Mahalanobis.Distance
130115	16233.809
239483	4204.829
178487	3807.244
151505	3436.276
10957	2877.559
220686	2729.211
85388	2592.572
78754	2285.229
137598	2269.037
16197	2184.955

Discussion

Using Mahalanobis Distance as our outlier detection measure, we found that observation 130,115 had the largest value (≈ 16,233), suggesting it is the most likely fraudulent transaction. In fact, the next largest Mahalanobis Distance was 4,204.83. The cutoff for inclusion among the top 250 potential fraud cases corresponded to a Mahalanobis Distance of 504.76. Across these 250 flagged transactions, the Mahalanobis Distances had a mean of 1,158.05 and a median of 1,078.99, indicating that while most of the suspicious transactions were tightly clustered, the mean was elevated by the extremity of observation 130,115.

Conclusion

The need for accurate fraud detection methods is critical as billions of dollars are lost to fraud in the US every year. Using a dataset of 250,000 financial transactions, we applied data transformations to reduce skewness and demonstrated that the transformed data more closely satisfied the assumptions of a multivariate normal distribution. Under this framework, we employed Mahalanobis Distance to identify outliers and flagged 250 transactions as the most likely to be fraudulent.

While we are confident in the validity of our results, future work could focus on additional transformations to further align the data with multivariate normality. Moreover, incorporating complementary approaches—such as PCA or ensemble methods that combine multiple outlier detection techniques—would provide a stronger and more robust basis for identifying potentially fraudulent activity.

References

Yeo, I.-K., & Johnson, R. A. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. Biometrika, 87(4), 954–959.

Mahalanobis, P.C. (1936) On the Generalized Distance in Statistics. Proceedings of the National Institute of Science of India, 2, 49-55.