Credit Card Fraud Detection with Multiple Correspondence Analysis and Linear Discriminant

Quan Tran

April 2024

Contents

1 Introduction

In the era of digital banking and e-commerce, credit card fraud has emerged as a significant global concern, impacting millions of consumers and businesses annually. The sophistication of fraudulent techniques has evolved, leveraging technology to exploit weaknesses in financial systems. Consequently, there is an urgent need for more advanced detection methods that can adapt to and anticipate the evolving tactics employed by fraudsters.

1.1 Research question and Motivation

This project is dedicated to developing a robust credit card fraud detection model using multivariate statistical methods. Our aim is to significantly reduce the incidence of fraud by improving the accuracy of fraud detection, thus minimizing financial losses and preserving consumer trust in credit and debit card transactions. By analyzing patterns and anomalies in transaction data, our system will not only identify existing fraudulent activities but also predict and prevent potential fraud scenarios before they occur.

1.2 Methodologies

The analysis involved using a range of statistical methods and techniques to detect fraud in the credit card transaction data. Summary statistics such as mean, median, and standard deviation were calculated, and Pearson's correlation coefficient and Chi-square tests were used to identify dependencies and multicollinearity. Various graphs and charts such as histograms, stacked-bar charts, and QQ plots were used to communicate the findings. Multiple Correspondence Analysis and Linear Discriminant Analysis methods were adopted due to its suitability. Justification for the use of these methods can be found under the section Method Selections.

1.3 Outline

The project will focus on several key objectives:

1. **Data set Description**: Describe of the data set to understand the meaning of the columns.
2. **Univariate analysis**: Explore each variable in a data set separately with summary statistics and visualization.
3. **Bivariate Analysis**: Explore the bivariate dependencies of the columns in the data set, augmented with visualization.
4. **Multivariate analysis**:
    - Justify for selection of the multivariate statistical methods.
    - Technical implementation of the method with R.
    - Results of the model.
5. **Critical evaluations**: Cevaet and report about possible sources of biases.
6. **Conclusion**

You can find the complete repository for the project here

2 Data and Analysis Problem

2.1 Dataset and Data Preprocessing

The Credit Card Fraud was obtained from the Kaggle website. The data set is sourced by some anonymous institute. It contains data 7 explanatory variables and 1 response variable, which is the classification whether a transaction is fraudulent for 1.000.000 transaction.

Finally, we extract only 3000 rows for training and validating with cross validation. While the training set is used for the explanatory data analysis and modeling, the test set is utilized for accuracy testing purposes.

Table 1: Descriptions of all columns in the dataset

Column	Description	Continuous/Binary
distance_from_home	The distance from home where the transaction happened.	Continuous
distance_from_last_transaction	The distance from last transaction happened.	Continuous
ratio_to_median_purchase_price	Ratio of purchased price transaction to median purchase price.	Continuous
repeat_retailer	Is the transaction happened from same retailer.	Binary
used_chip	Is the transaction through chip (credit card).	Binary
used_pin_number	Is the transaction happened by using PIN number.	Binary
online_order	Is the transaction an online order.	Binary
fraud	Is the transaction fraudulent.	Binary

2.2 Univariate Analysis

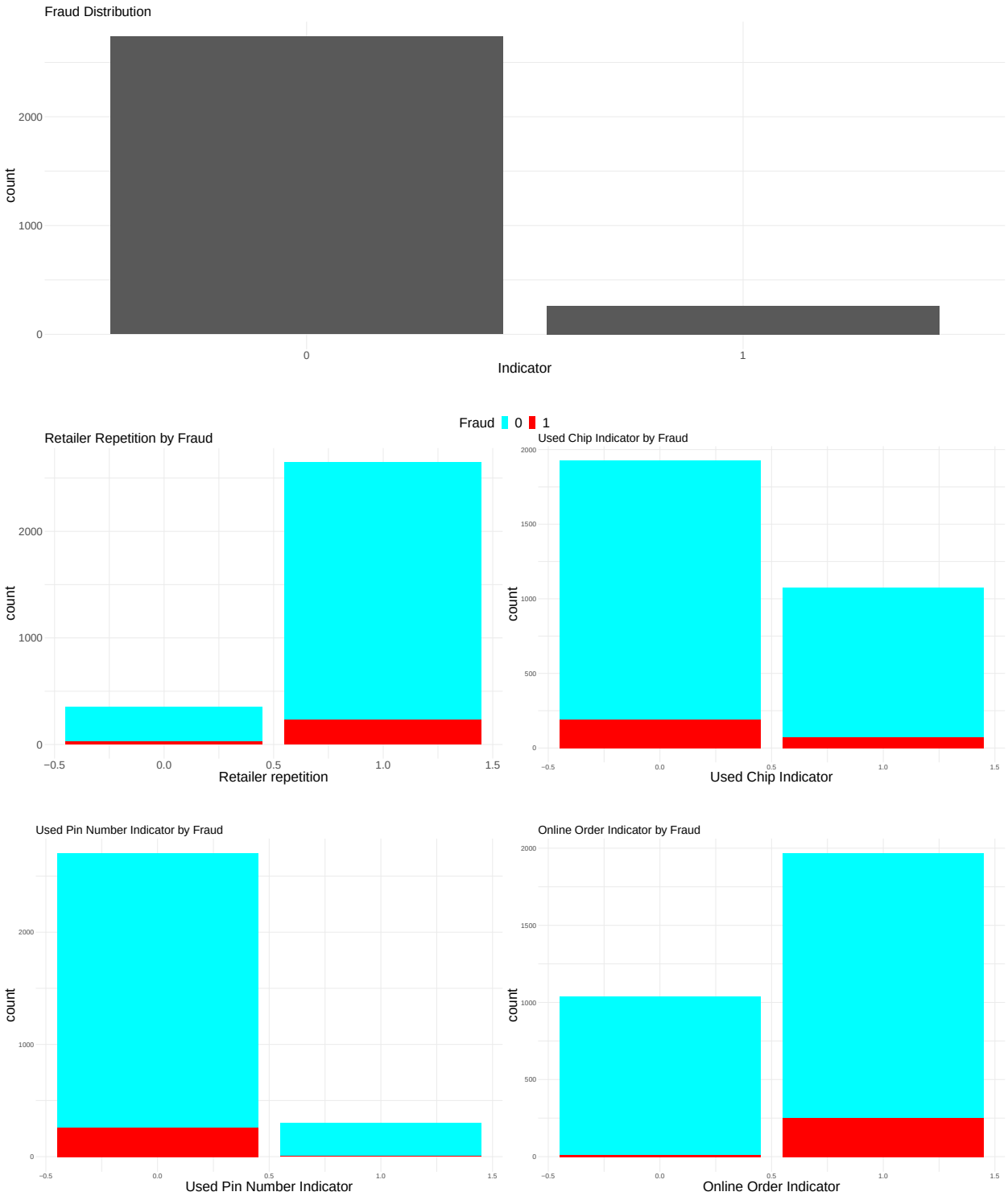
2.2.1 Summary Statistics

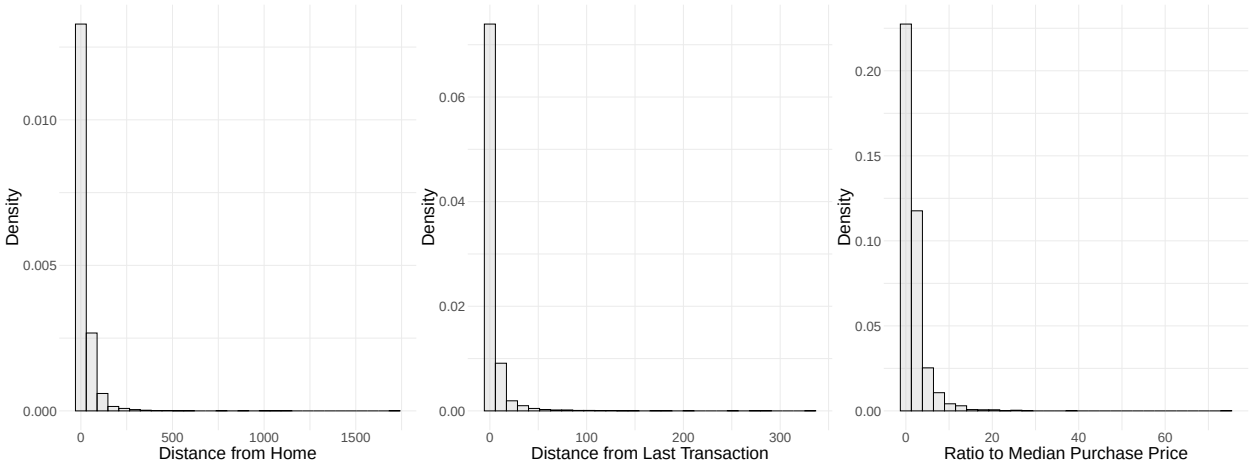
First of all, we will carry out some basic summary statistics calculations for every column in the data set.

```
distance_from_home  distance_from_last_transaction
Min.   :   0.0455   Min.   :  0.0030
1st Qu.:   3.9611   1st Qu.:  0.2859
Median :   9.9744   Median :  0.9665
Mean   :  25.6878   Mean   :  4.6107
3rd Qu.:  25.2605   3rd Qu.:  3.1799
Max.   :1712.1368   Max.   :331.0062
ratio_to_median_purchase_price repeat_retailer    used_chip
Min.   : 0.02627               Min.   :0.0000   Min.   :0.000
1st Qu.: 0.49965               1st Qu.:1.0000   1st Qu.:0.000
Median : 1.05014               Median :1.0000   Median :0.000
Mean   : 1.89205               Mean   :0.8827   Mean   :0.358
3rd Qu.: 2.09847               3rd Qu.:1.0000   3rd Qu.:1.000
Max.   :74.13623               Max.   :1.0000   Max.   :1.000
used_pin_number   online_order    fraud
Min.   :0.0000   Min.   :0.0000   0:2738
1st Qu.:0.0000   1st Qu.:0.0000   1: 262
Median :0.0000   Median :1.0000
Mean   :0.1003   Mean   :0.6547
3rd Qu.:0.0000   3rd Qu.:1.0000
Max.   :1.0000   Max.   :1.0000
```

2.2.2 Distribution Visualization

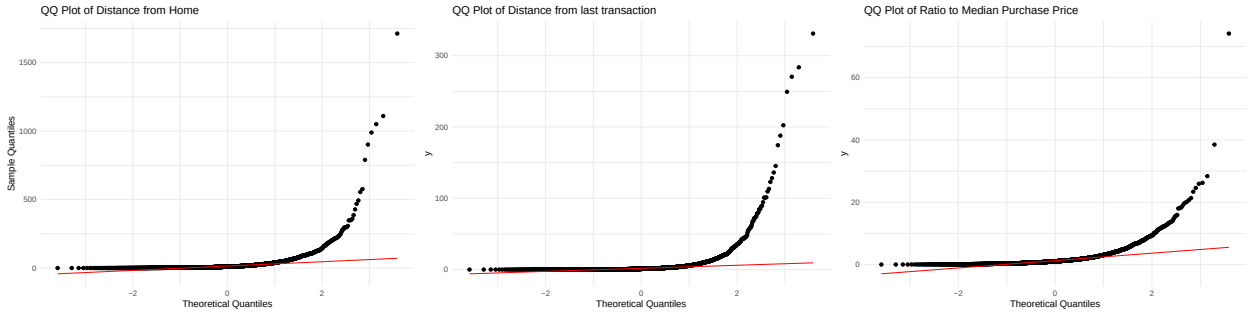
Plotting the histograms for the features that correspond to the response variable from the training dataset, we can gain insights into the distribution of the predictors and possible correlations between the explanatory variables and the response variable.



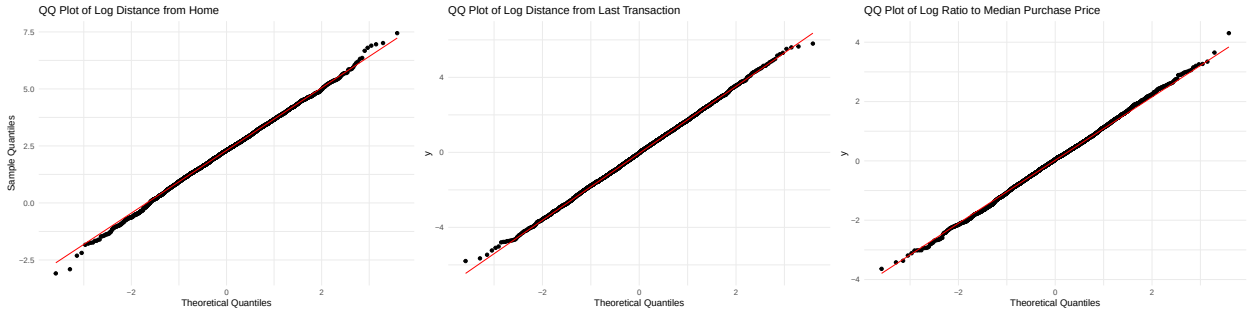


2.2.3 Normality Test

LDA assumes that the predictor variables for each class come from a multivariate normal distribution. This means that each class's predictor variables should roughly form a bell-shaped curve in a multidimensional space (where dimensions are equal to the number of predictors).In this part, we will plot QQ Plots to check for evidence of non-normality.



As notice from the plots, all three predictors are all right-skewed. However, normality can be achieved with log transformation since these columns are all positive.

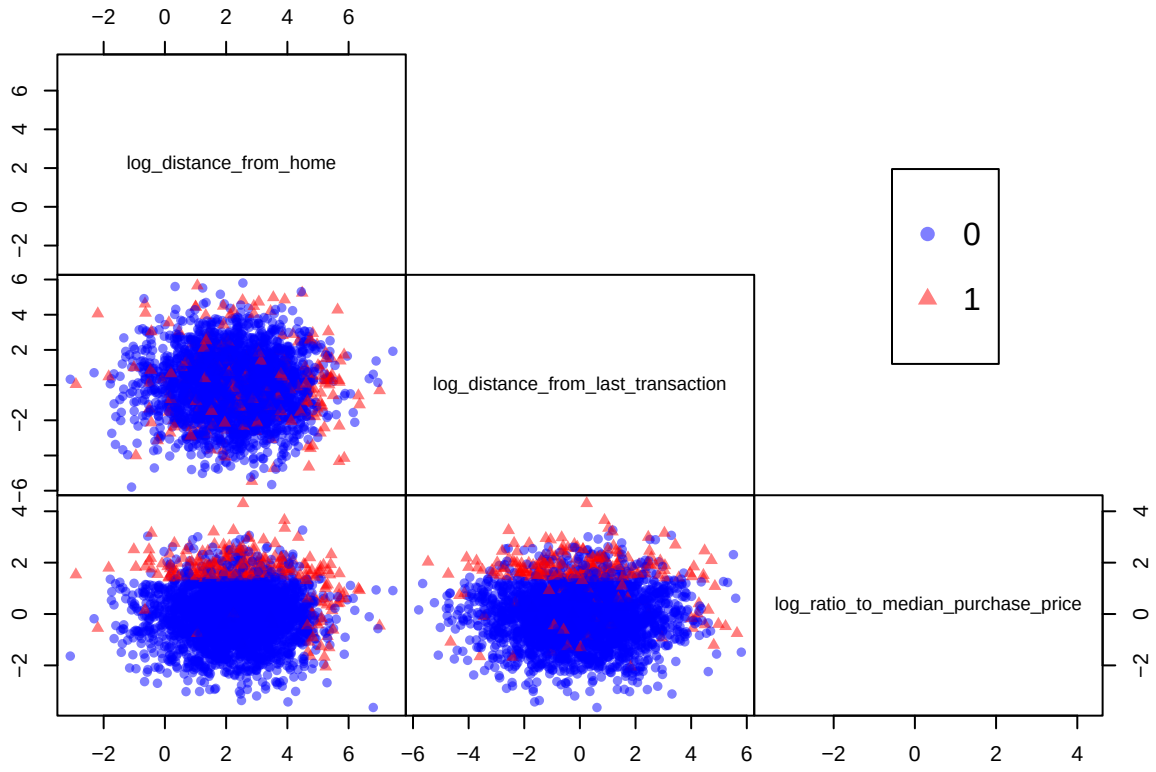


From the plots, it is safe to conclude that `Log Distance From Home`, `Log Distance from Last Transaction`, and `Log Ratio to Median Purchase Price` are all normally distributed since most of the points from each of the distribution lie on the lines.

2.3 Bivariate Analysis

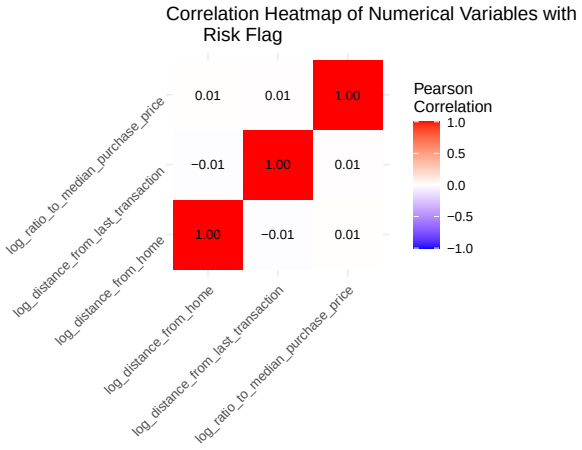
2.3.1 Scatter plots visualization

LDA assumes relatively low multicollinearity among predictors. High multicollinearity might exaggerate the estimated relationships among variables, affecting the stability of the coefficient estimates. Therefore, it is worth to check whether there are dependencies among the predictors.



2.3.2 Pearson's correlation coefficients

From the scatter plots, we can conclude that there are no correlations between the continuous variables `Log Distance From Home`, `Log Distance from Last Transaction`, and `Log Ratio to Median Purchase Price`. However, we can calculate the Pearson's correlation to confirm the uncorrelations.



As depicted in the heatmap, the correlations among the continuous predictors are approximately zero.

2.3.3 Chi-square tests

Table 2: Chi-square tests results

Pair	p-value	Result
repeat_retailer vs used_chip	0.3	Reject Null
repeat_retailer vs used_pin_num	0.17	Reject Null
repeat_retailer vs online_order	0.27	Reject Null
repeat_retailer vs fraud	0.67	Reject Null
used_chip vs used_pin_number	0.36	Reject Null
used_chip vs online_order	0.28	Reject Null
used_chip vs fraud	0.18	Reject Null
used_pin_number vs online_ord	0.79	Reject Null
used_pin_number vs fraud	1.09E-05	No Evidence
online_order vs fraud	3.42E-19	No Evidence

From the result table, we conclude that each of the pairs `used_pin_number and fraud` and `online_order and fraud` do not show a statistically significant association with the response variable with each other.

2.4 Multivariate Analysis

2.4.1 Method Selections

Two multivariate statistical methods will be used in this projects: Multiple Correspondence Analysis and Linear Discriminant Analysis.

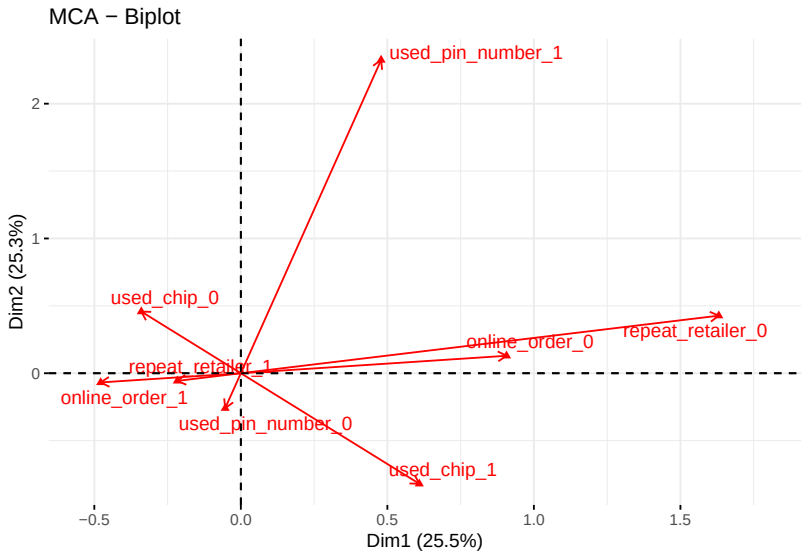
LDA is fundamentally a method for classification. It seeks to find a linear combination of features that characterizes or separates two or more classes of objects or events. The goal in credit card fraud detection is to distinguish between fraudulent and non-fraudulent transactions, making LDA an appropriate choice. Furthermore, LDA works by maximizing the ratio of between-class variance to the within-class variance in any particular data set, ensuring that the classes are as distinguishable as possible. Moreover, LDA is also generally less computationally intensive compared to methods like logistic regression or decision trees

when the primary goal is binary classification. This makes it an efficient choice given the large volume of transactions.

On the other hand, as LDA assumes the input dataset has a Gaussian distribution, we have to use MCA to transform set of binary variables into a smaller number of principal components (PCs). These components are continuous scores that represent the underlying patterns in the binary data.

2.4.2 MCA

After carrying out the Multiple Correspondence Analysis, we plot the MCA biplot to see the actractions/repulsions among the categories.



We get interpretation for this biplot:

- angle between modalities less than 90 degrees = attraction,
- angle between modalities more than 90 degrees = repulsion and
- angle between modalities 90 degrees = independent.

Then we use the all four resulted principal components as predictors for the LDA.

2.4.3 LDA

Before moving on with LDA, we standardize the continuous predictors to satisfy the Homogeneity of Variances and Covariances assumption of it.

```
card_extracted$log_distance_from_home <- scale(card_extracted$log_distance_from_home)[,1]
card_extracted$log_distance_from_last_transaction <- scale(card_extracted$log_distance_from_last_transa
card_extracted$log_ratio_to_median_purchase_price <- scale(card_extracted$log_ratio_to_median_purchase_
```

Then we utilize the components from MCA as predictors for the dependent variable.

```r
d_cv <- lda(fraud ~ log_distance_from_home + log_distance_from_last_transaction
            + log_ratio_to_median_purchase_price + PC1 + PC2 + PC3 + PC4,
            data = card_extracted, CV = TRUE)
```

```r
result <- data.frame(est = d_cv$class, truth = card_extracted$fraud)
```

The predictions are then compared to the true labels to compute the number of True Positive (TP), False Positive (FP), True Negative (NG), False Negative (NG).

```
Confusion Matrix:
```

```
                Actual False Actual True
Predicted False         2724         156
Predicted True            14         106
```

Then we will evaluate the classification accuracy of both models using some of the common metrics: Accuracy, Precision, Recall, and F1 Score.

- **Accuracy**: Accuracy measures how often a classifier makes the correct prediction. It is the ratio of the number of correct predictions to the total number of predictions.
- **Precision**: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It's a measure of a classifier's exactness. High precision relates to a low rate of false positives, and it is particularly important in cases where the cost of a false positive is high.
- **Specificity**: Specificity measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of non-creditworthy individuals who are correctly identified by the model).
- **Recall (or Sensitivity)**: Recall is the ratio of correctly predicted positive observations to all observations in the actual class. It's a measure of a classifier's completeness. High recall relates to a low rate of false negatives, and it is important in cases where the cost of a false negative is high.
- **F1 Score**: The F1 Score is the weighted average of Precision and Recall. It's a measure of a test's accuracy and considers both the precision and the recall. This is useful when seeking a balance between Precision and Recall, especially if there is an uneven class distribution.

```
Accuracy:    0.94
Precision:   0.88
Specificity: 0.99
Recall:      0.40
F1 Score:    0.55
```

The two metrics we would be focus on are Precision and Specificity. Since the cost false positivity is high (the loss of a credit card fraud might be very high), the Precision and Specificity metrics are appropriate since they put emphasis on classifier's exactness. The Precision and Specificity scores of this model are quite high (0.92 and 1.00, respectively), indicating that this model can be deployed for practical use.

To write the critical evaluations and conclusions for your project on credit card fraud detection using multivariate statistical methods, here are the sections carefully crafted to reflect thoughtful insight into the process and outcomes:

2.5 Critical Evaluations

Possible Sources of Bias and Limitations:

1. **Unbalanced Dataset:**
   - The dataset significantly leans towards non-fraudulent transactions, which is common in real-world scenarios but can introduce bias towards the majority class. This imbalance affects the training of the model, potentially leading to higher accuracy but poorer precision and recall for the minority class (fraudulent transactions).
   - **Mitigation Strategy:** Techniques like SMOTE (Synthetic Minority Over-sampling Technique), adjusted class weights, or anomaly detection methods could be considered to address this imbalance.
2. **Assumption of Normality:**
   - Linear Discriminant Analysis (LDA) assumes that the predictor variables are normally distributed within each class. The analysis highlighted non-normality which was adjusted using log transformations. However, this transformation might not perfectly normalize the data, potentially skewing the LDA results.
   - **Mitigation Strategy:** Continuous monitoring and validation of the normality assumption through more robust statistical tests or considering non-parametric methods if assumptions are consistently violated.
3. **Model Overfitting:**
   - Given the complexity of the methods used and the high dimensionality of the data, there is a risk of model overfitting where the model performs well on training data but less so on unseen data.

3 Conclusion

This project aimed to develop a robust credit card fraud detection system using a combination of Multiple Correspondence Analysis (MCA) and Linear Discriminant Analysis (LDA). The primary goal was to enhance the detection accuracy and thus reduce the incidence of fraud, safeguarding consumer transactions effectively.

- **Achievements:**
  - Successfully applied MCA to transform binary variables into principal components used in LDA, ensuring the dataset's suitability for the latter analysis.
  - The LDA model demonstrated high specificity and precision, indicating strong performance in identifying non-fraudulent transactions accurately and minimizing false positives, which is critical in financial contexts.
- **Performance Metrics:**
  - The model achieved a high overall accuracy and specificity, but the recall for fraudulent transactions was relatively low. This suggests that while the model is excellent at identifying legitimate transactions, it still misses a significant number of fraudulent cases.
- **Future Directions:**
  - Exploring additional feature engineering techniques and incorporating more diverse data sources to enrich the model's learning capacity.
- **Final Thoughts:**
  - Despite some limitations, the project establishes a solid foundation for a scalable and effective fraud detection system. Ongoing adjustments and enhancements based on emerging data and fraud techniques will be crucial to maintain and improve the system's efficacy.