Fraud Detection Statistical Analysis

Shahir Wakili

2025-12-29

Contents

1. Introduction

Fraud detection is a critical task in financial systems due to the high financial and reputational costs associated with fraudulent transactions. This analysis uses simulated mobile money transaction data to identify behavioural patterns associated with fraud.

From a statistical perspective, the key challenge is **class imbalance**, where fraudulent transactions represent only a small proportion of all observations. To address this, two complementary classification approaches are employed:

- **Logistic regression**, which provides interpretable parameter estimates and statistical inference.
- **Decision tree modelling**, which captures non-linear relationships and interaction effects that may not be well represented in parametric models.

Data Overview and Preprocessing

The dataset consists of transaction-level records, including transaction type, transaction amount, and account balance information before and after each transaction. These variables provide a foundation for modelling transactional behaviour associated with fraudulent activity.

Prior to analysis, data preprocessing and feature engineering were performed using **Python** to improve data quality and analytical usefulness. This process included the following steps:

- Removing non-informative identifier variables that do not contribute to predictive modelling.
- Creating new features to better capture transactional behaviour:
  - `deltaOrig`, representing the change in the sender's account balance.
  - `deltaDest`, representing the change in the recipient's account balance.
- Applying a logarithmic transformation to transaction amounts (`log_amount`) to reduce skewness and stabilise variance.
- Selecting a subset of relevant variables to support effective model training and interpretation.

After preprocessing, the cleaned dataset was exported and imported into **R** for further statistical analysis and model development.

2.1 Load and Inspect Data

```
## # A tibble: 6 x 5
##   type     log_amount deltaOrig deltaDest isFraud
##   <chr>         <dbl>     <dbl>     <dbl>   <dbl>
## 1 PAYMENT        9.19     9840.         0       0
## 2 PAYMENT        7.53     1864.         0       0
## 3 TRANSFER       5.20      181          0       1
## 4 CASH_OUT       5.20      181     -21182       1
## 5 PAYMENT        9.36    11668.         0       0
## 6 PAYMENT        8.96     7818.         0       0
```

```
##      type             log_amount       deltaOrig          deltaDest
##  Length:6362620     Min.   : 0.000   Min.   :-1915268   Min.   :-13060826
##  Class :character   1st Qu.: 9.502   1st Qu.:       0   1st Qu.:        0
##  Mode  :character   Median :11.224   Median :       0   Median :        0
##                     Mean   :10.841   Mean   :  -21231   Mean   :   124295
```

```
##                     3rd Qu.:12.249   3rd Qu.:   10150   3rd Qu.:   149105
##                     Max.   :18.342   Max.   :10000000   Max.   :105687839
##     isFraud
##  Min.   :0.000000
##  1st Qu.:0.000000
##  Median :0.000000
##  Mean   :0.001291
##  3rd Qu.:0.000000
##  Max.   :1.000000
```

Statistical purpose:
Initial inspection enables verification of variable types, identification of missing values, and detection of data quality issues. Summary statistics reveal strong skewness and scale differences, which are particularly relevant for fraud detection, as fraudulent transactions often exhibit extreme values and non-typical transaction behaviour that can influence model performance.

2.2 Variable Transformation

```
df$isFraud <- factor(df$isFraud, levels = c(0, 1))
df$type <- factor(df$type)

df$nameOrig <- NULL
df$nameDest <- NULL
```

Statistical purpose:
- The response variable (`isFraud`) is converted to a factor to enable classification modelling. - Transaction type is treated as a categorical predictor. - Identifier variables (`nameOrig`, `nameDest`) are removed because they do not carry predictive or statistical meaning and would introduce noise and overfitting.

The proportion of fraudulent transactions is examined to quantify class imbalance.

```
##
##          0          1
## 0.99870918 0.00129082
```

Interpretation: The proportion table shows that fraudulent transactions account for approximately 0.13% of all observations, confirming a severe class imbalance. This imbalance is characteristic of real-world fraud detection problems and necessitates the use of evaluation metrics such as sensitivity, precision, and AUC rather than overall accuracy.

3. Feature Engineering

```
##
##                   0       1
##   CASH_IN   1399284       0
##   CASH_OUT 2233384    4116
##   DEBIT      41432       0
##   PAYMENT  2151495       0
##   TRANSFER  528812    4097
```

Fraud is observed exclusively within the TRANSFER and CASH_OUT transaction types, while no fraudulent activity is recorded for CASH_IN, PAYMENT, or DEBIT transactions. This indicates that fraudulent behaviour in the dataset is concentrated in transactions involving the movement or withdrawal of funds. As a result, transaction type serves as a strong discriminative feature and justifies restricting subsequent analysis to these categories to improve model focus and performance.

```
##
##                   0       1
##   CASH_OUT 2233384    4116
##   TRANSFER  528812    4097
```

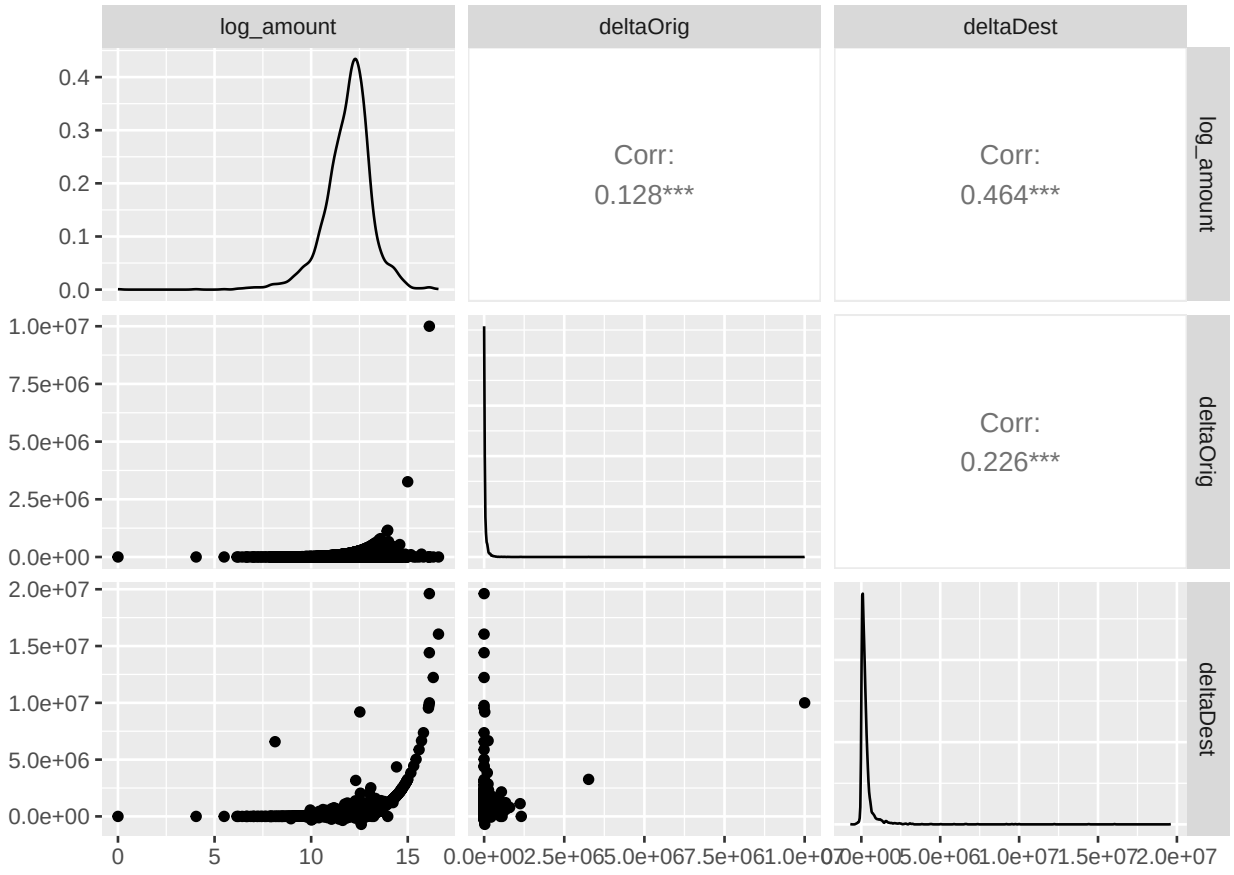
Statistical justification:
Including transaction types with zero recorded fraud can distort model estimation and artificially inflate classification accuracy without improving fraud detection capability. Therefore, the analysis is restricted to *TRANSFER* and *CASH_OUT* transactions, where fraudulent activity is observed, allowing the model to focus on meaningful discriminatory patterns.

4. Exploratory Data Analysis

4.1 Transaction Amount by Fraud Status

Statistical insight:
This comparison highlights distributional differences between fraudulent and legitimate transactions, fwraudulent transactions tend to involve higher transaction amounts and greater variability than non-fraudulent ones, indicating that transaction magnitude is a relevant predictor of fraud risk.

4.2 Pairwise Relationships



Statistical purpose:
Pairwise correlations show weak to moderate relationships between predictors, with correlations of approximately 0.13 (log_amount–deltaOrig), 0.46 (log_amount–deltaDest), and 0.23 (deltaOrig–deltaDest). These values indicate limited multicollinearity and support the inclusion of all engineered features in the model.

5. Train–Test Split

To evaluate model performance objectively, the dataset was partitioned into training and testing subsets using an 80–20 split. Stratified sampling was applied to preserve the original class distribution of the target variable (isFraud), ensuring that the severe class imbalance was consistently represented in both subsets. This approach reduces sampling bias and enables a more reliable assessment of model generalisation performance. Factor levels were also adjusted post-split to prevent unused categories from affecting model estimation. This separation ensures that model training and evaluation are conducted on independent data, providing an unbiased estimate of predictive performance.

```r
set.seed(42)
idx <- createDataPartition(df_stat$isFraud, p = 0.8, list = FALSE)
train <- df_stat[idx, ]
test <- df_stat[-idx, ]

train$type <- droplevels(train$type)
test$type <- droplevels(test$type)
```

6. Logistic Regression Model

```
##
## Call:
## glm(formula = isFraud ~ type + log_amount + deltaOrig + deltaDest,
##     family = binomial, data = train)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.716e+00  1.446e-01  -18.78   <2e-16 ***
## typeTRANSFER  1.503e+00  3.680e-02   40.85   <2e-16 ***
## log_amount   -3.586e-01  1.335e-02  -26.87   <2e-16 ***
## deltaOrig     1.426e-05  1.211e-07  117.73   <2e-16 ***
## deltaDest    -5.804e-06  7.996e-08  -72.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 89621  on 2216327  degrees of freedom
## Residual deviance: 41203  on 2216323  degrees of freedom
## AIC: 41213
##
## Number of Fisher Scoring iterations: 11
```

```
##       type log_amount  deltaOrig  deltaDest
##   1.089382   1.718326   3.014026   2.355800
```

Logistic Regression Model Diagnostics:
All predictors in the logistic regression model are statistically significant ($p < 0.001$), indicating that each variable contributes meaningfully to fraud prediction. The model achieves an AIC value of 41,213, reflecting a strong balance between model fit and complexity given the large sample size.

Variance Inflation Factor (VIF) values range from 1.09 to 3.01, which are well below commonly accepted thresholds (VIF < 5), indicating no evidence of problematic multicollinearity among predictors. This confirms that the engineered features provide independent information and that coefficient estimates are stable and reliable.

6.1 Logistic Regression Performance

```
## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction      0      1
##          0 552288    951
##          1    151    691
##
##                Accuracy : 0.998
##                  95% CI : (0.9979, 0.9981)
##     No Information Rate : 0.997
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5555
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.420828
##             Specificity : 0.999727
##          Pos Pred Value : 0.820665
##          Neg Pred Value : 0.998281
##              Prevalence : 0.002963
##          Detection Rate : 0.001247
##    Detection Prevalence : 0.001520
##       Balanced Accuracy : 0.710277
##
##        'Positive' Class : 1
##
```

Logistic Regression Performance Interpretation:
The model achieves an overall accuracy of 99.8%; however, given the extreme class imbalance, accuracy alone is not a reliable performance measure. The balanced accuracy of 0.71 provides a more meaningful assessment, indicating reasonable performance across both fraudulent and non-fraudulent classes.

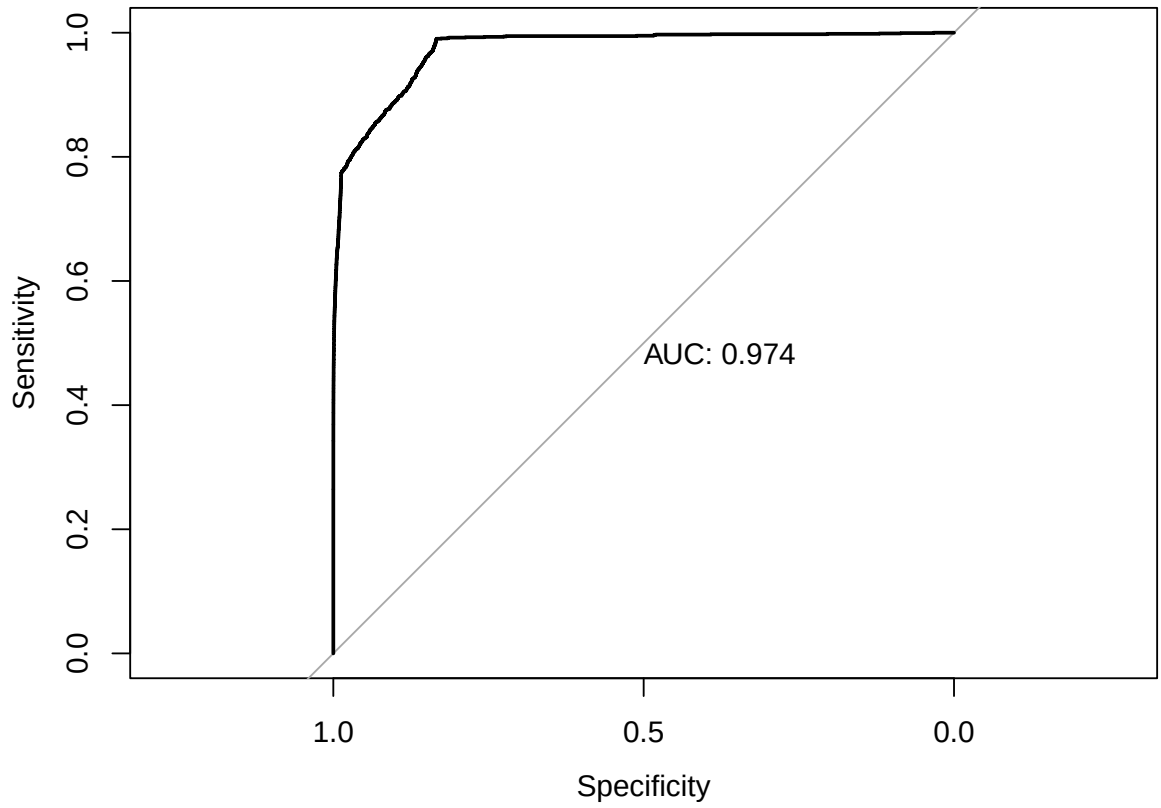
The model achieves a sensitivity of 0.42, meaning approximately 42% of fraudulent transactions are correctly identified. While this reflects moderate detection capability, it is expected in highly imbalanced fraud detection problems.

Specificity is very high (0.9997), indicating that legitimate transactions are almost always correctly classified. This contributes to a strong positive predictive value (0.82), meaning that when the model predicts fraud, it is correct in most cases.

6.2 Logistic ROC Curve

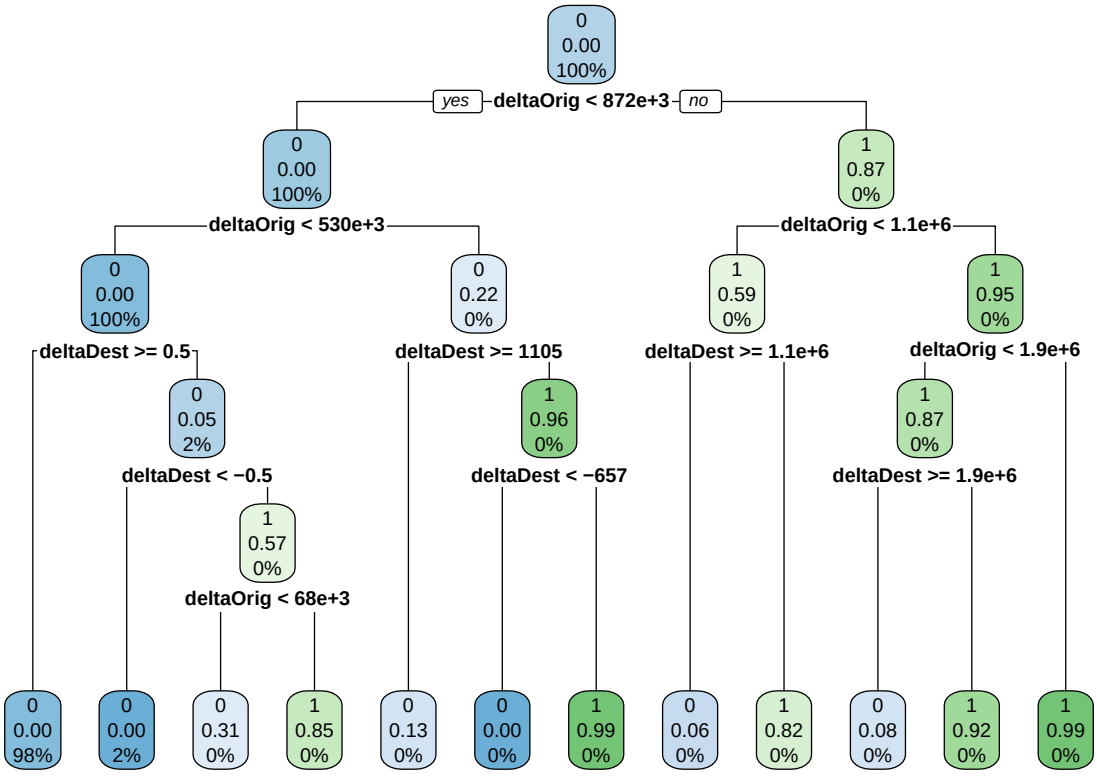
```
roc_glm <- roc(test$isFraud, prob_glm)
auc_glm <- auc(roc_glm)

plot(roc_glm, print.auc = TRUE)
```



ROC Curve Interpretation:
The ROC curve demonstrates strong discriminatory performance, with an Area Under the Curve (AUC) of 0.974. This indicates that the model has a very high ability to distinguish between fraudulent and non-fraudulent transactions across all classification thresholds.

7. Decision Tree Model



Statistical motivation:
The decision tree identifies balance change variables as the primary drivers of fraud detection, with large reductions in the origin account balance strongly associated with fraudulent transactions. Several terminal nodes exhibit high predicted fraud probabilities (exceeding 90%), indicating strong model confidence when extreme balance movements occur. In contrast, transactions with minimal balance changes are consistently classified as legitimate. This demonstrates the model's ability to capture non-linear threshold effects that are not well represented by linear models, making it particularly effective for identifying high-risk transaction patterns in highly imbalanced fraud data.

7.1 Decision Tree Performance

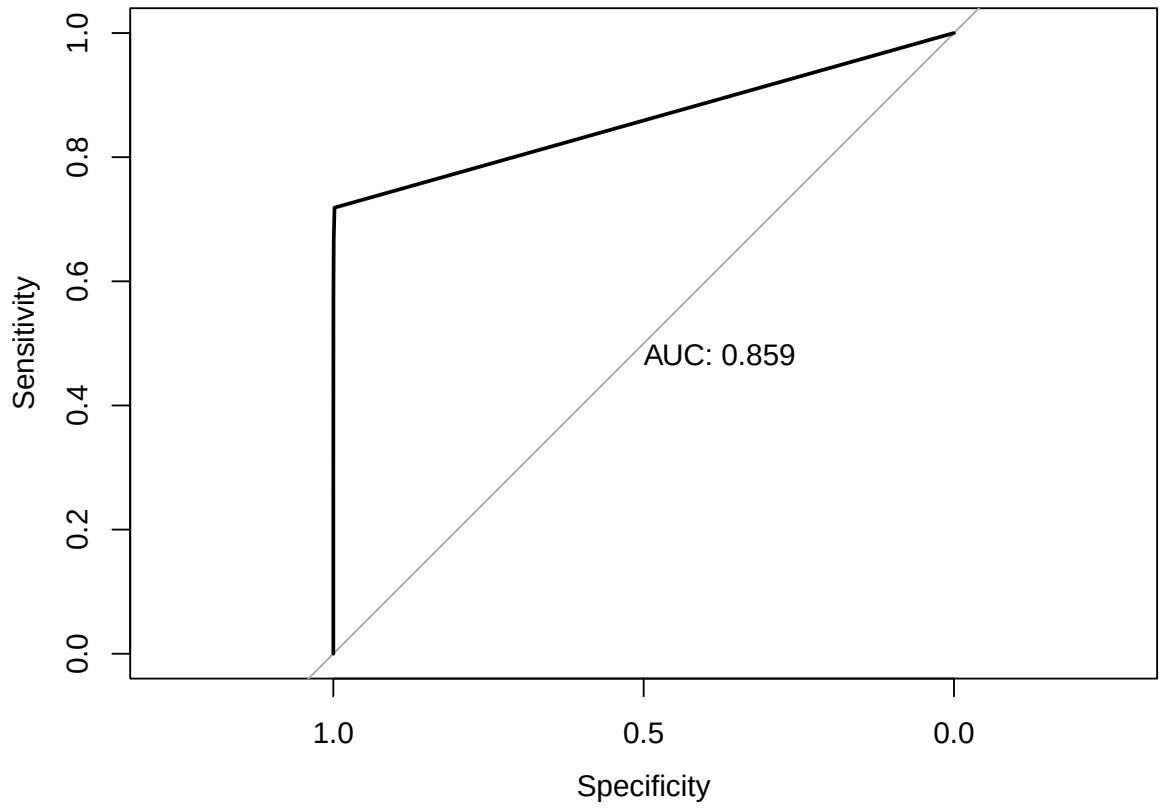
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 552331    686
##          1    108    956
##
##                Accuracy : 0.9986
##                  95% CI : (0.9985, 0.9987)
##     No Information Rate : 0.997
```

```
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7059
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.582217
##             Specificity : 0.999805
##          Pos Pred Value : 0.898496
##          Neg Pred Value : 0.998760
##              Prevalence : 0.002963
##          Detection Rate : 0.001725
##    Detection Prevalence : 0.001920
##       Balanced Accuracy : 0.791011
##
##        'Positive' Class : 1
##
```

Decision Tree Performance

The decision tree achieved a high overall accuracy of 99.86%, with a balanced accuracy of 0.79, indicating improved performance across both classes compared to the logistic model. The model demonstrates strong fraud detection capability, with a sensitivity of 0.58 and a high precision of 0.89, meaning most flagged transactions are truly fraudulent. Specificity remains very high (0.9998), showing minimal false positives. Overall, the decision tree provides stronger detection of fraudulent activity while maintaining reliable classification performance.

7.2 Decision Tree ROC Curve



Decision Tree ROC Curve Interpretation

The decision tree achieves an AUC of 0.859, indicating strong discriminatory ability. Although slightly lower than the logistic regression model, it effectively captures non-linear fraud patterns and provides meaningful classification performance.

8. Performance Comparison of Classification Models

```
##                 Model       AUC Sensitivity Specificity Precision
## 1 Logistic Regression 0.9735268   0.4208283   0.9997267 0.8206651
## 2       Decision Tree 0.8589792   0.5822168   0.9998045 0.8984962
##   Balanced_Accuracy
## 1         0.7102775
## 2         0.7910107
```

Comparative Model Evaluation:
The logistic regression model achieves a higher AUC (0.974), indicating stronger overall discrimination between fraudulent and non-fraudulent transactions. However, the decision tree demonstrates superior sensitivity (0.58 vs 0.42) and balanced accuracy (0.79 vs 0.71), indicating better performance in identifying fraudulent cases under class imbalance. Both models exhibit very high specificity (>0.99), meaning legitimate transactions are rarely misclassified.

Overall, while logistic regression provides stronger global ranking performance, the decision tree offers improved fraud detection capability and more balanced classification, making it more suitable when identifying fraudulent transactions is the primary objective.

9. Conclusion

This analysis explored fraud detection using both logistic regression and decision tree models on transactional data characterised by significant class imbalance. Through feature engineering, particularly the use of balance change variables and log-transformed transaction amounts, meaningful patterns associated with fraudulent behaviour were successfully captured. Exploratory analysis demonstrated that fraud is strongly concentrated in specific transaction types and is associated with large, abnormal balance movements.

Model evaluation showed that while logistic regression achieved superior overall discrimination (AUC = 0.974), the decision tree provided stronger performance in identifying fraudulent transactions, as reflected by its higher sensitivity and balanced accuracy. This highlights an important trade-off between global predictive performance and effective fraud detection in imbalanced datasets. The decision tree's ability to model non-linear relationships and capture threshold-based risk patterns makes it particularly valuable in operational fraud detection contexts.

Overall, the results demonstrate that combining interpretable statistical models with flexible machine learning approaches provides a robust framework for fraud detection. Future work could explore threshold optimisation, cost-sensitive learning, or ensemble methods to further improve detection performance while minimising false positives.