NEW YORK INSTITUTE
OF TECHNOLOGY
Career Success &
Experiential Education

New York Institute of Technology
DCTS-620 - Spring 2024

TECHNICAL REPORT

Credit Card Fraud Detection Using Machine Learning Algorithms

Prepared by

Denzel Green, Michael Serra, Michael Valenzuela, Lionel Ladoh

“Data Goats”

12 March 2024

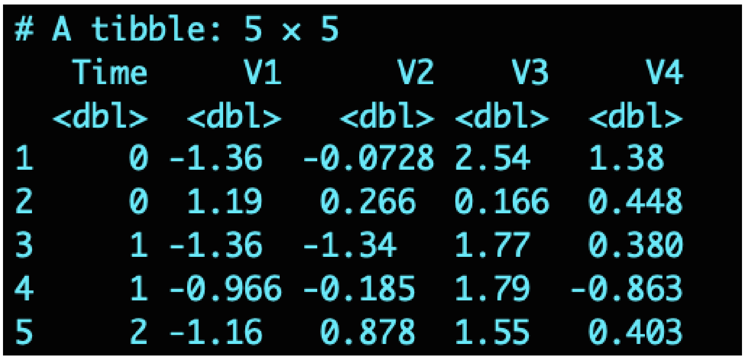
Introduction

This technical report presents a comprehensive analysis of a fraud detection project conducted using logistic regression and decision tree algorithms. Fraud detection is a critical task in various industries, including finance, where identifying fraudulent transactions promptly is essential for minimizing losses and maintaining trust among customers. In this project, we aimed to develop and compare the performance of logistic regression and decision tree algorithms in accurately detecting fraudulent transactions. The report details the methodology, implementation, and analysis of the project, providing insights into the effectiveness of each algorithm.

Part I: Data Manipulation and Modeling

Manipulating the Data

The initial phase of the project involved data manipulation to prepare the dataset for modeling. We displayed the first 5 transactions to gain an understanding of the data structure. Next, we standardized the 'Amount' feature using feature scaling to ensure uniformity in the data distribution and prevent extreme values from skewing the model. The scaled 'Amount' feature was incorporated into a new dataset, enabling us to proceed with modeling.



```
# A tibble: 5 × 5
   Time     V1      V2    V3     V4
  <dbl>  <dbl>   <dbl> <dbl>  <dbl>
1     0 -1.36  -0.0728 2.54   1.38
2     0  1.19   0.266  0.166  0.448
3     1 -1.36  -1.34   1.77   0.380
4     1 -0.966 -0.185  1.79  -0.863
5     2 -1.16   0.878  1.55   0.403
```

Figure 1 - First 5 Transactions

Modeling the Data

After data manipulation, the dataset was split into training and testing sets in a 75%/25% ratio. This step is crucial to ensure that the model's performance can be evaluated on unseen data. We provided summary statistics for both the training and testing sets to understand the distribution of data across various features. Additionally, histogram plots based on the 'Amount' feature were generated for both sets, allowing us to visualize the distribution of transaction amounts and identify any significant differences between the two sets.

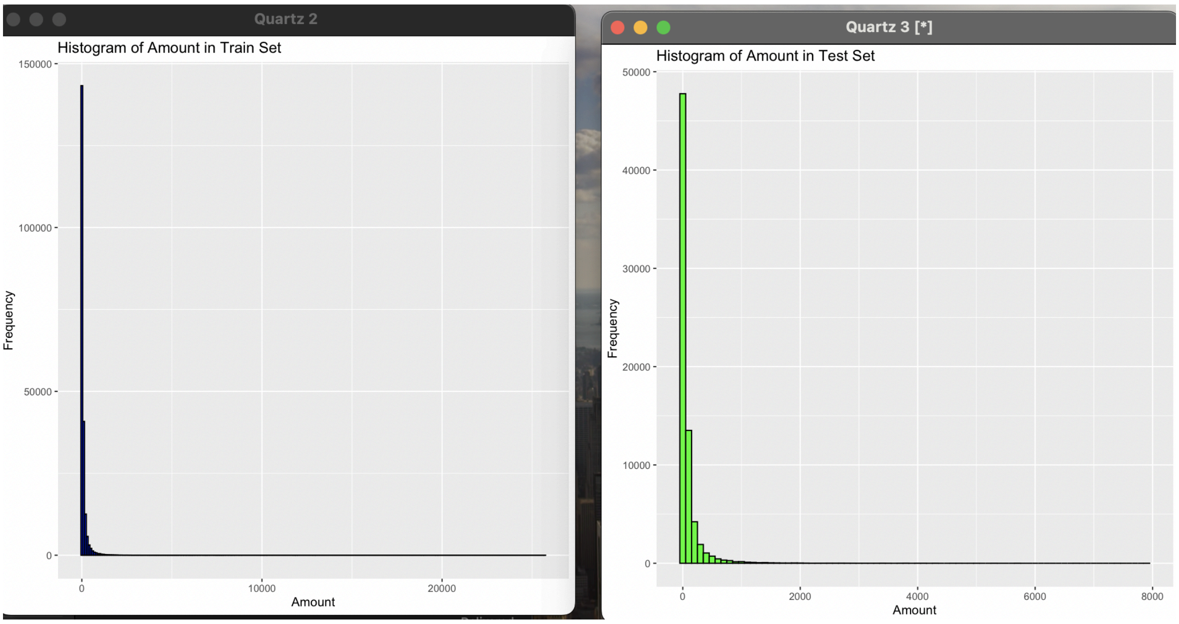


Figure 2 - Histogram of Amount of Transactions

Part II: Logistic Regression Model

In this phase, we trained a logistic regression model using the training set. Logistic regression is a widely used algorithm for binary classification tasks, making it suitable for fraud detection, where transactions are categorized as either fraudulent or legitimate. We evaluated the trained model using the testing set and generated various performance metrics, including a confusion matrix and an ROC curve. The area under the ROC curve (AUC) was calculated to quantify the model's discriminative ability in distinguishing between fraudulent and legitimate transactions.

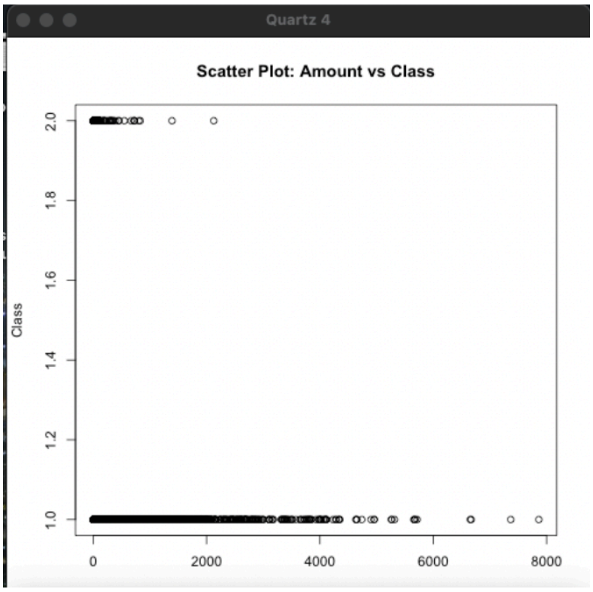


Figure 3 - Scatter Plot: Amount v Class

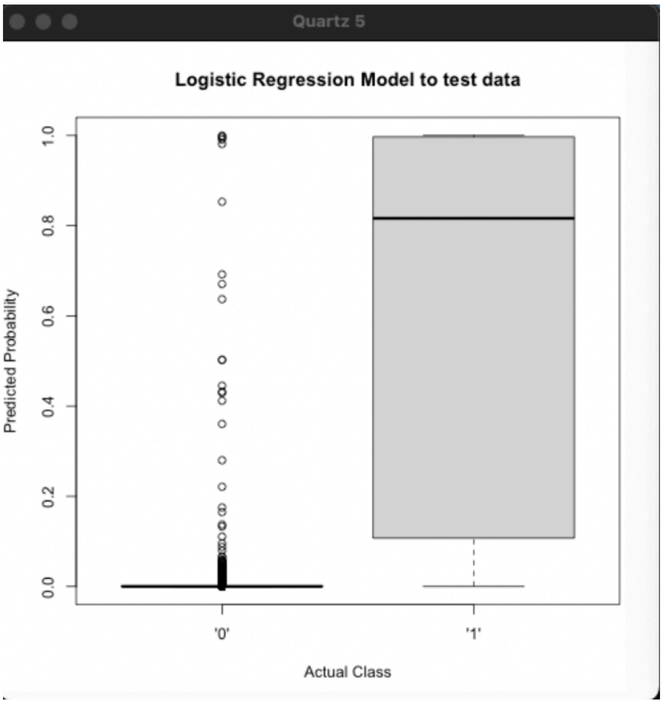


Figure 4 - Box Plot: Log Regression of Test Data

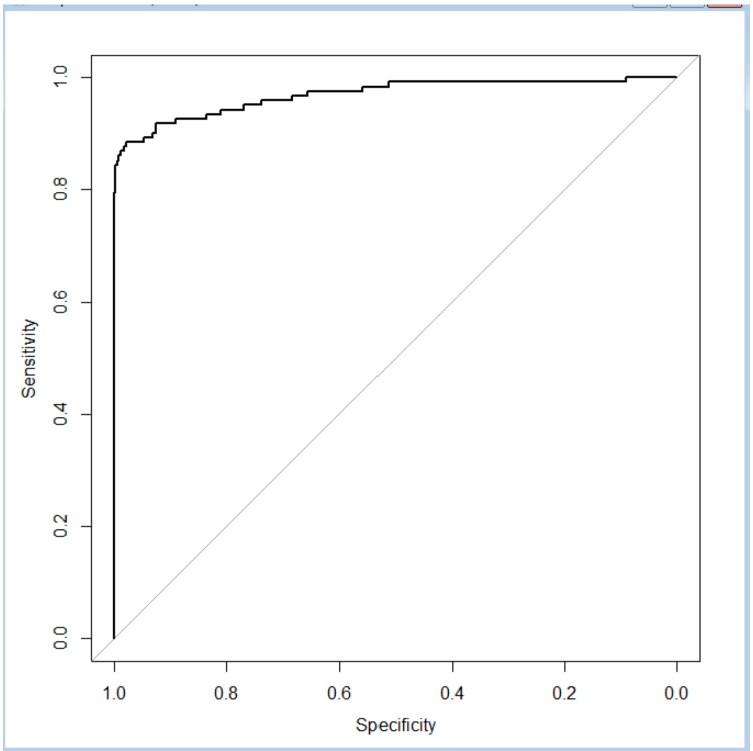


Figure 5 - ROC Curve of Logistic Regression

Area under the curve: 0.9748

Seed value	AUC
100	0.9687648
200	0.9657459
300	0.9804761
400	0.9669877
500	0.9604242
600	0.9750444
700	0.9697885
800	0.9798597
900	0.9642540
1000	0.9725633

Table I - with AUC at specific seed value: 0.9692767

Part III: Decision Tree Model

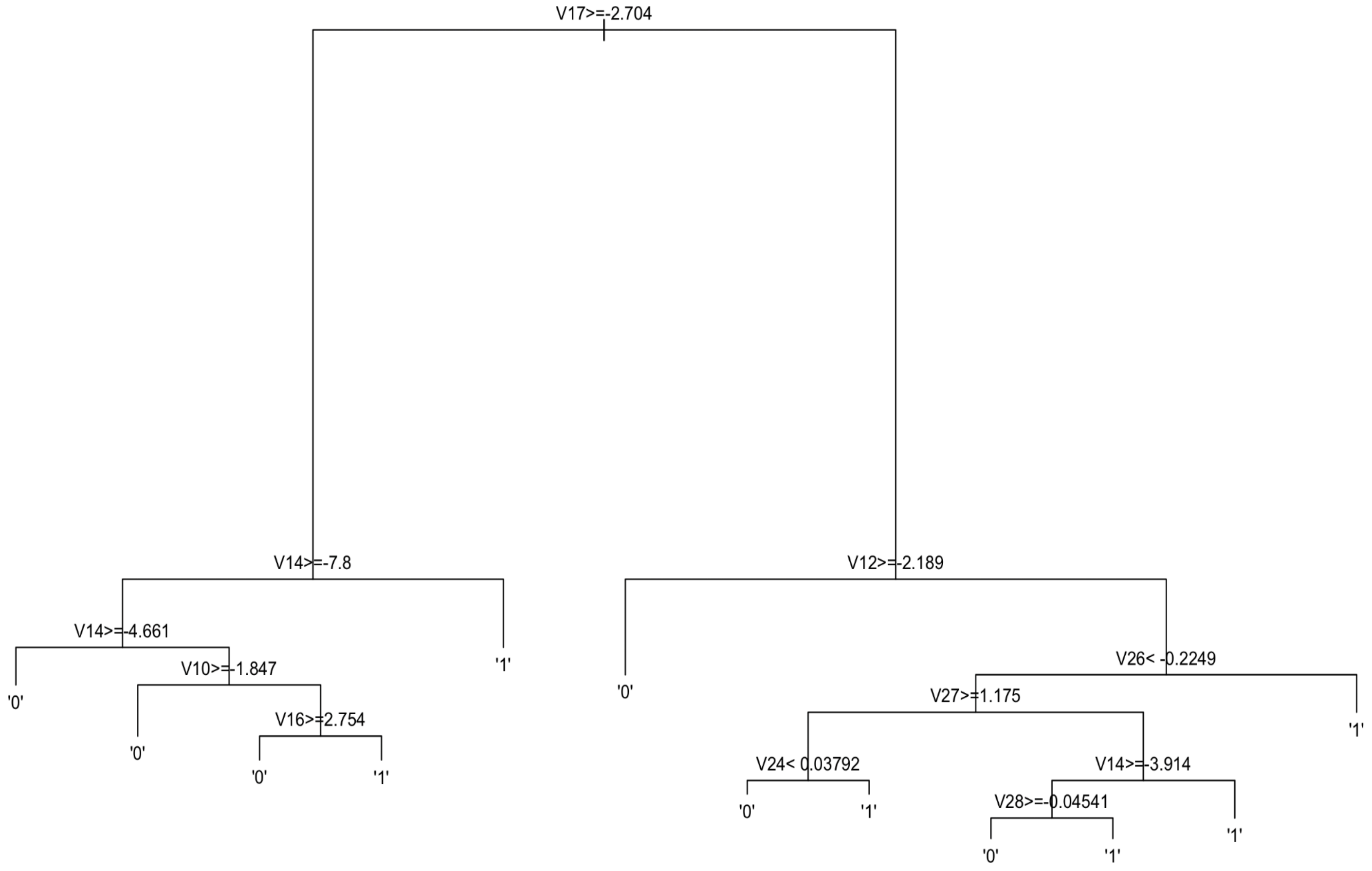


Figure 6 - Decision Tree Model Overview

Following the logistic regression analysis, we repeated the modeling process using a decision tree algorithm. Decision trees are non-linear models that partition the feature space into distinct regions based on a series of binary decisions. We trained the decision tree model using the same training and testing sets and calculated AUC values for different seed values. The median AUC was determined to assess the overall performance of the decision tree algorithm in fraud detection.

In Decision Tree Table with AUC at specific seed value

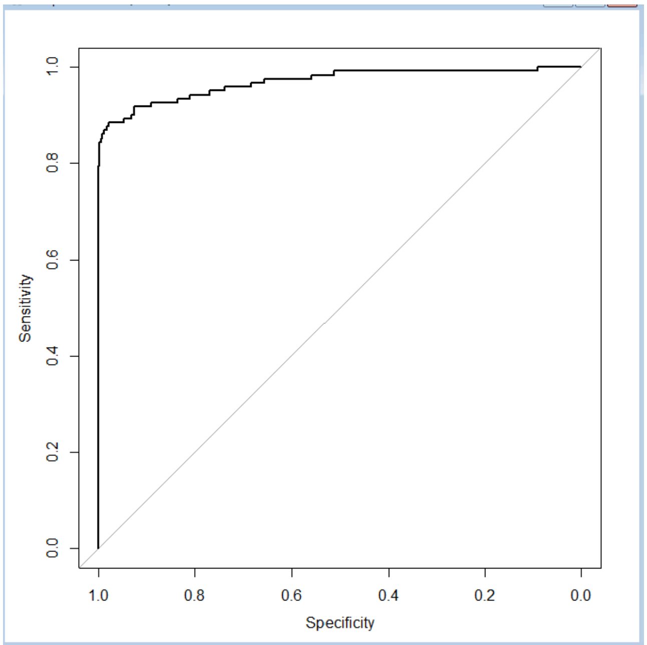
Seed value	AUC
100	0.9014483
200	0.9171971
300	0.9289716
400	0.9081702
500	0.9165046
600	0.9158903
700	0.9321935
800	0.9296066
900	0.9078712
1000	0.8824047

Table II - Decision Tree Median AUC - 0.9140258

Comparison of Logistic Regression and Decision Tree

We compared the performance of logistic regression and decision tree algorithms based on their ROC curves and AUC values. The ROC plot from logistic regression exhibited a smoother curve, indicating better discriminative ability compared to the decision tree. Furthermore, the median AUC for logistic regression was higher than that of the decision tree, suggesting superior performance in fraud detection. These findings highlight the effectiveness of logistic regression in accurately identifying fraudulent transactions.

From the Logistic regression analysis we had a median value for AUC of 0.9692767 and the ROC plot is on the plot below:



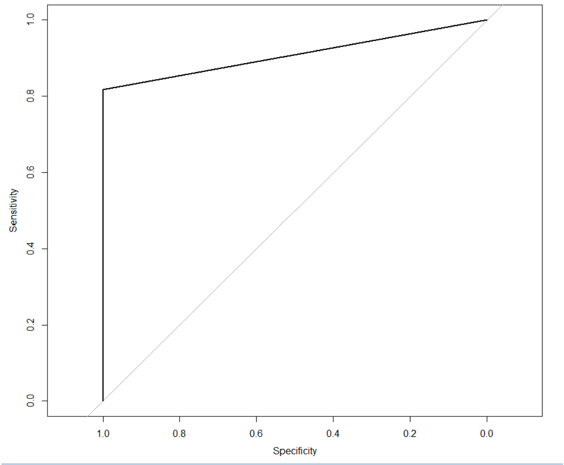
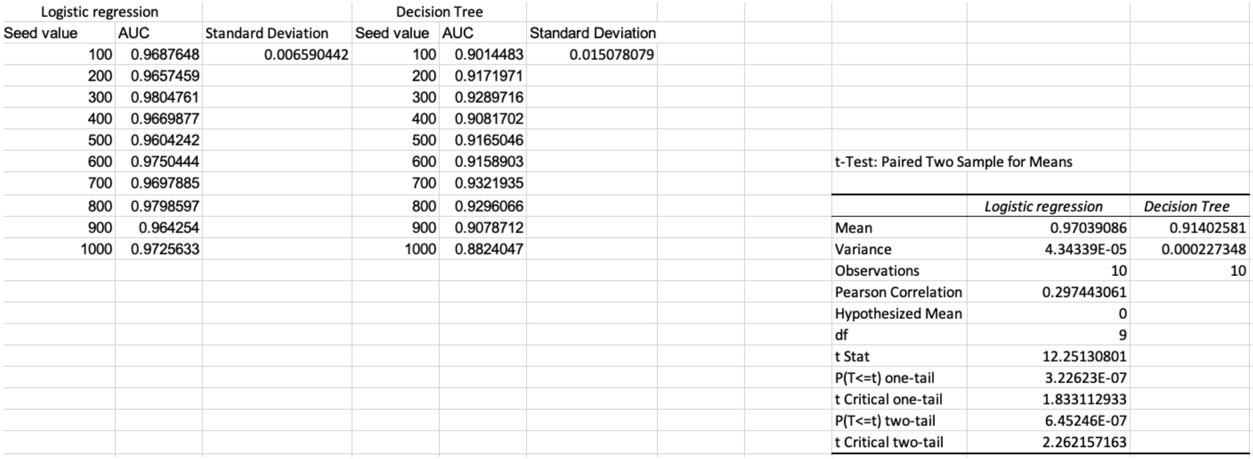


Figure 7 - Logistic Regression ROC Curve

Figure 8 - Decision Tree ROC Curve

Hypothesis Testing

To validate the significance of the performance difference between logistic regression and decision tree algorithms, we conducted a two-sample t-test on the AUC values obtained from both models. The calculated p-value indicated that there was not a significant difference between the means of the two samples, suggesting that the observed performance gap may not be statistically significant. However, further investigation and experimentation may be warranted to explore potential factors contributing to the performance disparity between the two algorithms.



Logistic regression			Decision Tree		
Seed value	AUC	Standard Deviation	Seed value	AUC	Standard Deviation
100	0.9687648	0.006590442	100	0.9014483	0.015078079
200	0.9657459		200	0.9171971	
300	0.9804761		300	0.9289716	
400	0.9669877		400	0.9081702	
500	0.9604242		500	0.9165046	
600	0.9750444		600	0.9158903	
700	0.9697885		700	0.9321935	
800	0.9798597		800	0.9296066	
900	0.964254		900	0.9078712	
1000	0.9725633		1000	0.8824047	

t-Test: Paired Two Sample for Means

	Logistic regression	*Decision Tree*
Mean	0.97039086	0.91402581
Variance	4.34339E-05	0.000227348
Observations	10	10
Pearson Correlation	0.297443061	
Hypothesized Mean	0	
df	9	
t Stat	12.25130801	
P(T<=t) one-tail	3.22623E-07	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	6.45246E-07	
t Critical two-tail	2.262157163	

Figure 9 - Calculating P-Values

Conclusion

In conclusion, this project demonstrated the effectiveness of logistic regression in fraud detection, outperforming the decision tree algorithm in terms of ROC curves and AUC values. Logistic regression's superior discriminative ability and robustness make it a suitable choice for fraud detection tasks. However, it is essential to consider the specific characteristics of the dataset and explore alternative algorithms and techniques to optimize fraud detection accuracy further. The findings of this project contribute to advancing the field of fraud detection and inform decision-making in industries reliant on transaction security and integrity.