School of Computing
FACULTY OF ENGINEERING

UNIVERSITY OF LEEDS

Credit Card Fraud Detection

Shengxuan Ji

Submitted in accordance with the requirements for the degree of MSc Advanced Computer Science (Data Analytics)

2023

The candidate confirms that the following have been submitted.

Items	Format	Recipient(s) and Date
Deliverable 1	Report	SSO (DD/MM/YY)
Deliverable 2	Code for the whole project	SSO (DD/MM/YY)

Type of project: Exploratory Software

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

(Signature of Student) ______________________

Summary

<Concise statement of the problem you intended to solve and main achievements (no more than one A4 page)>

Acknowledgements

<The page should contain any acknowledgements to those who have assisted with your work. Where you have worked as part of a team, you should, where appropriate, reference to any contribution made by other to the project.>
Note that it is not acceptable to solicit assistance on 'proof reading' which is defined as the "the systematic checking and identification of errors in spelling, punctuation, grammar and sentence construction, formatting and layout in the test";
see http://www.leeds.ac.uk/gat/documents/policy/Proof-reading-policy.pdf.

Contents

Chapter 1

Introduction

1.1 Problem Statement

Credit card transaction fraud is when fraudsters steal consumer credit information through various means and then impersonate the cardholder to make purchases. It is a very common financial issue. When fraud occurs, the cardholder's financial security is threatened, and merchants risk losing merchandise and being punished. In this age of rapid information technology development, cash transactions are becoming less and less frequent, and transaction fraud is constantly occurring, posing a significant threat to people's financial security.

When a transaction takes place, various attributes are recorded simultaneously (such as transaction time, location, credit card number, amount, etc.) at the service provider's end [CreditCardFraudDetection4]. This data can be used to check for fraudulent transactions and even interrupt suspicious transactions before they are completed and contact the cardholder.

The main problem is the human inability to instantly recognise fraudulent transactions. As a result, the implementation of machine learning algorithms become important for the analysis of transactional data and detecting such fraudulent activities. In this context, a range of algorithmic models is utilised. And these models' performance is then systematically evaluated through comparative analysis.

1.2 Aim and Objectives

1.2.1 Aim

This project aims to:

1. Identify the variables that have the highest predictive power for fraudulence, meaning the variables that have the most significant impact on the outcome.

2. Compare various models and determine which model has the highest accuracy, as well as understanding the reasons for its superior performance.

1.2.2 Objectives

The objectives of this project are to:

1. Perform data pre-processing, including using Synthetic Minority Oversampling Technique (SMOTE) on credit card transactions data, and conduct a visualisation analysis and Z-test to detect any potential patterns and determine the most impactful variables.

2. Utilise both supervised and unsupervised machine learning algorithms for detecting transactional fraud, including logistic regression, random forest, Support Vector Machine (SVM), Isolation Forest, and Local Outlier Factor. Additional algorithms, such as Naive Bayes Classifier and Autoencoder, may be utilised if time permits.

3. Conduct several performance evaluation metrics, including accuracy, precision, recall, and ROC curve analysis, to assess the effectiveness of the models. This will be followed by a detailed analysis of the underlying factors contributing to the superior performance of the selected models. If a better approach is discovered during project implementation, it will be adopted.

4. Present a summary of the key points related to Anomaly Detection, highlight the most important takeaways from the analysis, and propose improvement measures.

1.3 Risk Assessment

Risk	**Likelihood**	**Impact**	**Mitigation plan**
No original data provided because confidence	Cell 2	Cell 3	Cell 4
Software bugs	Cell 6	Cell 7	Cell 8
No enough time to finish full project	Medium	High	Cell 12
File Missing	Low	High	Use the Internet hosting service to keep fil

Table 1.1: Table Title

Chapter 2

Background

2.1 Machine learning for fraud detection

2.1.1 Supervised Learning

2.1.2 Unsupervised Learning

2.2 Importance of Predictive Variables in Fraud Detection

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. [1]

Chapter 3

Methodology

3.1 Tools

Remember why and why not

3.2 Dataset

3.3 logistic regression

logistic regression

3.4 random forest

3.5 Support Vector Machine(SVM)

3.6 Naive Bayes Classifier

3.7 Isolation Forest

3.8 Local Outlier Factor

3.9 Auto-encoder

3.10 Evaluation Metrics

3.10.1 Accuracy

3.10.2 precision

3.10.3 recall

3.10.4 ROC Curve

Chapter 4

Data Preparation

4.1 Data Pre-processing

4.1.1 SMOTE

4.1.2 Z-test

4.2 Data Visualisation

Chapter 5

Implementation

5.1 Experiment

5.1.1 Model 1 - logistic regression

5.1.2 Model 2 - random forest

5.1.3 Model 3 - Support Vector Machine(SVM)

5.1.4 Model 4 - Naive Bayes Classifier

5.1.5 Model 5 - Isolation Forest

5.1.6 Model 6 - Local Outlier Factor

5.1.7 Model 7 - Auto-encoder

5.2 Results Discussion

Chapter 6

Evaluation

6.1 Comparison between diff modules

6.1.1 Accuracy

6.1.2 precision

6.1.3 recall

6.1.4 ROC Curve

Chapter 7

Conclusion

7.1 Overview

7.2 Challenges

7.3 Limitations

7.4 Future Work

7.5 Personal Reflection

References

[1] J. Smith and M. Johnson. Research methods in ai. *Journal of Artificial Intelligence*, 23(2):34–56, 2009.

Appendices

Appendix A

External Material

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Appendix B

Ethical Issues Addressed