**UNIVERSITY OF BRIGHTON**

**MODULE TUTOR: -**

**MODULE CODE: - CI603**

**MODULE TITLE: - - Data mining**

**STUDENT ID: - 21839630**

**SUBMISSION DATE: - 23/05/2024**

**Topic:- Credit Card Fraud Data Detection with Data mining**

Credit Card Fraud Data Detection with Data mining.

Contents

[1.Introduction 2](#_Toc167352295)

[Objective 2](#_Toc167352296)

[Dataset Description 2](#_Toc167352297)

[Project Scope 2](#_Toc167352298)

[2.Data Exploration and Preprocessing 2](#_Toc167352299)

[Initial Data Analysis 2](#_Toc167352300)

[Handling Missing Values and Duplicates 3](#_Toc167352301)

[Correlation Analysis 3](#_Toc167352302)

[Feature Scaling 3](#_Toc167352303)

[Handling Imbalanced Data 4](#_Toc167352304)

[3.Dimensionality Reduction 4](#_Toc167352305)

[Theoretical Background 4](#_Toc167352306)

[Principal Component Analysis (PCA) 4](#_Toc167352307)

[Linear Discriminant Analysis (LDA) 4](#_Toc167352308)

[Application: 4](#_Toc167352309)

[4.Classification Models 5](#_Toc167352310)

[Training and Evaluation 5](#_Toc167352311)

[Performance Metrics: 5](#_Toc167352312)

[Comparison and Relevance to Data Mining: 5](#_Toc167352313)

[5.Anomaly Detection 6](#_Toc167352314)

[Techniques Used 6](#_Toc167352315)

[Isolation Forest: 6](#_Toc167352316)

[Autoencoders: 6](#_Toc167352317)

[Evaluation: 6](#_Toc167352318)

[6.Critical Evaluation and Discussion 7](#_Toc167352319)

[Evaluation of Techniques: 7](#_Toc167352320)

[Dimensionality Reduction 7](#_Toc167352321)

[Classification Models 7](#_Toc167352322)

[Anomaly Detection 7](#_Toc167352323)

[Advantages and Disadvantages 8](#_Toc167352324)

[7.Conclusion 8](#_Toc167352325)

[Reference 8](#_Toc167352326)

# 1.Introduction

## Objective

The objective of this data mining project is to implement different data mining techniques to detect the fraud transaction of the credit card dataset. In today’s increase of the volume in online transactions credit card fraud has become the most notable concern for the financial institutions and users are getting adversely getting affected by the fraud transactions. To solve this problem effective methods have been employed in this project to stop the financial lose and protect the user data (Bhattacharyya et al., 2011).

## Dataset Description

The data set used in this project consists of credit card transactions over the period. Each of these transactions is categorized by the set features which are derived from the principal Component Analysis (PCA), this is done to anomalies to protect the data of the cardholder information and a binary target variable indicating whether the transaction is fraudulent. The dataset is very complex containing 2,84,807 transactions with 31 columns. The dataset is highly imbalanced with fraud transactions making up to only 0.172% of the total data.

## Project Scope

This project performs multiple stages of data mining techniques which are data exploration, data cleaning and preprocessing, dimensionality reduction using the principal component analysis and Linear Discriminant analysis. Classification of models such as Logistic regression, Decision Tree and Random Forest and the anomaly detection is done by the Insolation Forest and Autoencoders. Each of these techniques will be evaluated by the metrics for effectiveness of detecting the fraud transactions.

# 2.Data Exploration and Preprocessing

## Initial Data Analysis

The initial exploration of the dataset was involved in calculating of the statistic and visualizing the distribution. The dataset was highly imbalanced because of the fraud transaction being very low with the number of 473 and non-fraudulent data being 283253. Visualization, such as histograms and box plots were created to understand distribution and variability of the features.

A graph of a distribution of time

Description automatically generated

**Figure 1Distribution of time.**

In the amount feature it exhibits a right-skewed distribution, which indicates that few transactions were with very high amounts and in the time feature, it represents the seconds elapsed between each transaction and the first transaction. Which shows no relationship with the pattern of the fraud.

## Handling Missing Values and Duplicates

There are no missing values in the dataset. However, the duplicate entries were identified, there were of 1081 duplicate entries, all these entries were removed using the function **duplicated(),** which resulted in clean data set with 283726 unique transactions. Handling duplicated transactions was crucial because the skew the data, so they were removed and ensured the reliability of subsequent analysis. (Kwak and Kim, 2017).

## Correlation Analysis

Correlation analysis was performed to understand the relationships between features and their relationship with the target variable. A heatmap of the correlation matrix was created to visualize these relationships.

A graph of a heatmap

Description automatically generated

**Figure 2.Correlation heatmap.**

Most features had low correlations with each other, indicating minimal multicollinearity. The 'Amount' feature had a low correlation with the target variable, suggesting it alone is not a strong predictor of fraud.

## Feature Scaling

The scaling of features was performed using StandardScaler to standardize the features. This process ensured that each feature had a mean of 0 or a standard deviation of 1. This helps to standardizing the data to improve the performance of many classifying models by ensuring that all the features contribute equally to the distance calculations used in these model (Patro and Sahu, 2015).

By doing this it results in more stable and reliable performance for the classifying models like logistic regression and Random Forest and the transformed features has mean close to 0 and standard deviation close to 1.

## Handling Imbalanced Data

Since the dataset was highly imbalanced, this can cause models to become biased towards predicting the majority classes into non-fraudulent transactions. The Synthetic Minority Over-Sampling Technique (SMOTE) was employed to address it. SMOTE works by creating synthetic samples for the minority class (fraudulent transactions) by interpolating between existing minority class samples. This technique helps to balance the dataset, enabling models to learn and predict the minority class more effectively (Chawla et al., 2002).

After applying the SMOTE, the dataset became balanced, with an equal number of fraudulent and non-fraudulent transactions. This balance is crucial for training effectively with the classification models. Chawla et al., 2002).

# 3.Dimensionality Reduction

Dimensionality reduction approaches were used to enhance our models' functionality and effectiveness. By keeping the most crucial data and lowering the feature count, this benefits the dataset.

## Theoretical Background

Two popular dimensionality reduction techniques were used, which are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

### Principal Component Analysis (PCA)

PCA is an unsupervised approach that transforms the data into a new coordinate system where the greatest variances by any projection of the data come to lie on the first coordinates, which are called principal components. PCA is particularly useful when dealing with high-dimensional data, since it minimizes the dimensionality by mapping the data onto a subspace that is lower dimensional while maintaining the greatest amount of variance. (Salih Hasan and Abdulazeez, 2021).

### Linear Discriminant Analysis (LDA)

LDA is a supervised feature extraction and dimensionality reduction technique, meaning it does not focus on capturing the maximum variance inside a set of data, as is the case with PCA. In contrast, LDA aims to find a feature subspace with good class separability. Through this, data is projected into a low-dimensional space such that there exists good class separability to avoid overfitting, yet it remains computationally efficient (Chen et al., 2019).

### Application:

**PCA:** We applied PCA to our standardized dataset. The cumulative explained variance plot indicated that 13 principal components were sufficient to retain 95% of the variance in the data. By reducing the dataset from its original 30 features to 13 principal components, we achieved a more manageable and computationally efficient dataset.

A graph with a line

Description automatically generated

**Figure 3.PCA: Explained Variance vs. Number of Components.**

A graph showing a diagram of a person's body

Description automatically generated

**Figure 4.PCA: First 2 Principal Component**

The first two principal components captured a significant portion of the variance, allowing for a visual representation of the data. The scatter plot of the first two principal components showed some separation between the fraudulent and non-fraudulent transactions, although the classes were not entirely distinct.

**LDA:** We applied LDA to the dataset, focusing on maximizing the separation between the fraudulent and non-fraudulent transactions. Given that LDA is a supervised method, it reduced the data to a single component that best discriminates between the two classes.

A graph with a line in the center

Description automatically generated

**Figure 5.LDA: Linear Discriminant 1**

The transformation resulted in a single component that provided better separation between the classes compared to the first two PCA components. This improved separation is expected to enhance the performance of the classification models.

# 4.Classification Models

To detect fraud Transaction, we have used the three classifications models which are Logistic Regression, Decision Tree, and Random Forest. Each model was evaluated using both PCA and LDA transformed data to determine which dimensionality reduction technique provided the best performance (Pranckevičius and Marcinkevičius, 2017).

## Training and Evaluation

Each model was trained and tested using both PCA and LDA transformed datasets. The dataset was first standardized and then balanced using SMOTE. Later, the dataset was split into training (80%) and testing (20%) sets. Performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, were calculated to evaluate to assess how well the models identified fraudulent transactions.

## Performance Metrics:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Dimensionality Reduction** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| Logistic Regression | PCA | 0.934900 | 0.972309 | 0.895741 | 0.932455 | 0.984858 |
| Logistic Regression | LDA | 0.930522 | 0.958038 | 0.900966 | 0.928626 | 0.976099 |
| Decision Tree | PCA | 0.996196 | 0.996147 | 0.996270 | 0.996209 | 0.996196 |
| Decision Tree | LDA | 0.896242 | 0.898464 | 0.894228 | 0.896341 | 0.896459 |
| Random Forest | PCA | 0.999267 | 0.999384 | 0.999156 | 0.999270 | 0.999996 |
| Random Forest | LDA | 0.896295 | 0.898221 | 0.894650 | 0.896432 | 0.955276 |

## Comparison and Relevance to Data Mining

The PCA version of the Random Forest model outperformed all the other models and their configurations. It has the best accuracy, precision, recall, F1 score, and ROC-AUC among all the models, thereby proving itself to be the best model for identifying fraudulent transactions in this dataset. Logistic Regression was also quite good, especially in performed with PCA. The Decision Tree model proved to have high accuracy when in performed with PCA but rather poor concentration with LDA.

These models describe the practical implementation of these data mining techniques in the identification of fraud. Both PCA and LDA worked effectively in reducing dimensionality, which affects computational efficiency and model performance. As can be seen from the better results of Random Forest with PCA, strong classification algorithms are of importance.

# 5.Anomaly Detection

In addition to classification models, we used anomaly detection techniques to identify fraudulent transactions. Anomaly detection is particularly useful in fraud detection scenarios because fraudulent transactions are rare and often deviate significantly from normal patterns.

## Techniques Used

### Isolation Forest

This technique isolates observations by randomly selecting a feature and splitting the data. It works on the principle that anomalies are few and different, and therefore they are easier to isolate. Isolation Forest is efficient and well-suited for high-dimensional datasets(Calheiros et al., 2017).

Insolation Forest was applied by standardized dataset. This was trained to find any anomalies by isolating observations. The contamination parameter was set to the proportion of fraudulent transactions in the dataset to help the model learn the distribution of anomalies.

### Autoencoders:

Autoencoders are neural networks used to learn efficient coding of input data. They consist of an encoder that compresses the data and a decoder that reconstructs it. Anomalies are detected based on the reconstruction error; higher errors indicate anomalies (Torabi et al., 2023).

The autoencoder was trained on the standardized dataset, minimizing the reconstruction error. The model was designed to reconstruct the input data, and anomalies were identified based on their high reconstruction error compared to normal transactions.

## Evaluation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| Isolation Forest | 0. 697543 | 0. 699065 | 0. 697233 | 0. 698148 | 0. 697544 |
| Autoencoder | 0. 503186 | 0. 984127 | 0. 009817 | 0. 019440 | 0. 504829 |

Isolation Forest significantly outperformed the autoencoder in terms of accuracy, precision, recall, and F1-score. The autoencoder, despite having a high precision, had extremely low recall and F1-score, indicating poor overall performance in detecting anomalies.

Since, Isolation Forest demonstrated better performance for anomaly detection, providing a practical choice for real-time fraud detection systems. Although Autoencoders are capable of modelling complex data distributions, their effectiveness in this scenario was limited.

These techniques complemented the classification models by providing an additional layer of security, particularly effective in identifying new or rare fraud patterns that classification models might miss. Integrating anomaly detection with classification models achieved a more comprehensive and robust fraud detection system (Calheiros et al., 2017)

# 6.Critical Evaluation and Discussion

## Evaluation of Techniques:

The techniques used in the project have individual strength and limitations. A contrasting performance shall provide us an insight into its practical utility toward fraud detection.

### Dimensionality Reduction

**PCA:** PCA was helpful to reduce the dimensionality of the dataset because it retains 95% of the variance and will reduce the file size and computational time in applying models in the classification. However, PCA is an unsupervised technique, and components obtained cannot always be the best discriminative components for class separation(Salih Hasan and Abdulazeez, 2021).

**LDA:** LDA provided a low-dimensional space that maximizes the class separability and is particularly useful for classification tasks. While in general, LDA performed well, with difficulties in case of nonlinear class boundaries. However, its usage requires the assumption that classes should share the same covariance matrix.

### Classification Models

**Logistic Regression:** Logistic regression performed decently with both PCA and LDA, and slightly better with PCA. It is an interpretable straightforward model and simple to train, but the simple model structure can fail to capture patterns in complicated, nonlinear datasets(Pranckevičius and Marcinkevičius, 2017).

**Decision Tree:** PCA Decision Tree model performed well with PCA but not for LDA. It is intuitive and easy to interpret the model in the form of a set of rules and visualization but holds poor performances in high dimensions because it is likely to be overfitted(Pranckevičius and Marcinkevičius, 2017).

**Random Forest:** The Random Forest with respect to the Random Forest with respect to the PCA test outperformed most of the models. Since, it is an ensemble model and it inherits the strengths of the multiple decision trees present in the model, making it really accurate and robust against overfitting. In turn, it is one of the most computationally costly and less interpretable simple models.

### Anomaly Detection

**Isolation Forest** Anomaly detection was performed more accurately and more precisely with Isolation Forest. It is computationally efficient and scales very well with big data. But, due to its randomness feature, it sometimes gives variant results.

**Autoencoders** were quite a remarkable unsupervised learning approach in the context of anomaly detection because they could model highly non-linear data distribution and thus represent the underlying data with high-dimensional complexities. Although the approach is efficient, it is computationally demanding and requires careful hyperparameter tuning.

## Advantages and Disadvantages

**Advantages** PCA and Random Forest have both given the best performances recorded, which means robust feature selection, coupled with powerful classification algorithms. Added security with the Isolation Forest is amongst other fraud patterns, which are rare and novel (Calheiros et al., 2017).

**Disadvantages** The data set is highly imbalanced, and it is not easy to handle, even with techniques like SMOTE. The assumption previously mentioned in the LDA, about equal variance for all the variables and its linear approach, makes it less efficient compared to PCA. The computational demands of Random Forest and Autoencoders could be a different negative attribute for a real-time application(Chawla et al., 2002).

# 

# 7.Conclusion

This project has successfully applied various data mining techniques to detect fraud credit card transactions. The combination of PCA for dimensionality reduction and Random Forest for classification achieved the highest performance, highlighting the importance of feature selection and powerful algorithms. Isolation Forest provided effective anomaly detection, adding an extra layer of security. These findings demonstrate that integrating dimensionality reduction and classification models can significantly enhance fraud detection capabilities. Future work could focus on incorporating advanced deep learning techniques and real-time data updates to further improve the system's accuracy and adaptability.

# Reference

Bhattacharyya, S., Jha, S., Tharakunnel, K. and Westland, J.C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3), pp.602–613. doi:<https://doi.org/10.1016/j.dss.2010.08.008>.

Calheiros, R.N., Ramamohanarao, K., Buyya, R., Leckie, C. and Versteeg, S. (2017). On the effectiveness of isolation-based anomaly detection in cloud data centers. *Concurrency and Computation: Practice and Experience*, 29(18), p.e4169. doi:<https://doi.org/10.1002/cpe.4169>.

Chan, P.K., Fan, W., Prodromidis, A.L. and Stolfo, S.J. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems*, 14(6), pp.67–74. doi:<https://doi.org/10.1109/5254.809570>.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(16), pp.321–357. doi:<https://doi.org/10.1613/jair.953>.

Chen, D.-W., Miao, R., Yang, W.-Q., Liang, Y., Chen, H.-H., Huang, L., Deng, C.-J. and Han, N. (2019). A Feature Extraction Method Based on Differential Entropy and Linear Discriminant Analysis for Emotion Recognition. *Sensors*, 19(7), p.1631. doi:<https://doi.org/10.3390/s19071631>.

Fraud Detection of Credit Card using Data Mining Techniques. (2019). *International Journal of Innovative Technology and Exploring Engineering*, 8(12), pp.4410–4413. doi:<https://doi.org/10.35940/ijitee.l3952.1081219>.

Kwak, S.K. and Kim, J.H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, 70(4), pp.407–411. doi:<https://doi.org/10.4097/kjae.2017.70.4.407>.

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M.-S. and Zeineddine, H. (2019). An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. *IEEE Access*, 7, pp.93010–93022. doi:<https://doi.org/10.1109/access.2019.2927266>.

Patro, S.Gopal.K. and sahu, K.K. (2015). Normalization: A Preprocessing Stage. *IARJSET*, pp.20–22. doi:<https://doi.org/10.17148/iarjset.2015.2305>.

Pranckevičius, T. and Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing*, 5(2). doi:<https://doi.org/10.22364/bjmc.2017.5.2.05>.

Salih Hasan, B.M. and Abdulazeez, A.M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *Journal of Soft Computing and Data Mining*, 2(1). doi:<https://doi.org/10.30880/jscdm.2021.02.01.003>.

Torabi, H., Mirtaheri, S.L. and Greco, S. (2023). Practical autoencoder based anomaly detection by using vector reconstruction error. *Cybersecurity*, 6(1). doi:<https://doi.org/10.1186/s42400-022-00134-9>.