A Comprehensive Investigation into Anomaly Detection in Credit Card Transactions Using Autoencoder Variants

A PREPRINT

Aaditya Awadhiya
Department of Computer Science
University of Stuttgart
Stuttgart, 70569
st185630@stud.uni-stuttgart.de

Sabrina Caspary
Department of Computer Science
University of Stuttgart
Stuttgart, 70569
scaspary@gmx.de

February 8, 2024

ABSTRACT

The area of financial transactions, credit card fraud poses a constant threat, demanding sophisticated detection methodologies. This research embarks on an extensive exploration of the application of autoencoders, specifically variational autoencoders (VAEs), sparse autoencoders, normal autoencoders, and denoising autoencoders, for credit card fraud detection. Leveraging a Kaggle dataset that encompasses labeled instances of both normal and fraudulent credit card transactions, we delve into two distinct scenarios: Case 1, where only normal transactions contribute to the training set, and Case 2, featuring a mixed dataset with both normal and anomalous instances. A novel two-step training model, coupled with meticulous data cleaning, is introduced for the mixed data scenario. This paper aims to provide a profound understanding of autoencoder functionality in anomaly detection and meticulously compares their efficacy in diverse scenarios.

1 Introduction

1.1 Background

The persistent threat of credit card fraud in the financial sector necessitates the continuous evolution of detection methods. Anomaly detection, focusing on the identification of patterns or instances significantly deviating from the norm, stands out as a crucial aspect in fraud detection. Autoencoders, a class of neural networks, present a promising avenue due to their ability to capture intricate patterns. This study seeks to explore their potential in detecting anomalies within credit card transactions.

1.2 Objectives

This research endeavors to compare the performance of different autoencoder variants—VAEs, sparse autoencoders, normal autoencoders, and denoising autoencoders—in credit card fraud detection. The evaluation is conducted under two scenarios: Case 1, where unsupervised learning is applied using only normal data for training, and Case 2, where a mixed dataset with both normal and anomalous instances is considered. Additionally, the paper provides a comprehensive overview of anomaly detection, explains the suitability of autoencoders for this task, and delves into the unique advantages offered by each autoencoder variant.

2 Anomaly Detection

2.1 Definition

Anomaly detection, synonymous with outlier detection, involves the identification of instances in a dataset that substantially deviate from the norm or expected behavior. In the context of credit card fraud detection, anomalies represent transactions that deviate from typical, legitimate patterns, indicating potential fraudulent activity.

2.2 Importance in Credit Card Fraud Detection:

Swift detection of anomalies is imperative to minimize financial losses and protect users from fraudulent transactions. Traditional methods often struggle to adapt to evolving fraud patterns, underscoring the significance of anomaly detection in modern fraud prevention systems.

3 Autoencoders for Anomaly Detection

3.1 Working Principle

Autoencoders, designed for unsupervised learning, consist of an encoder and a decoder. The encoder encodes essential features of the input data into a latent space, and the decoder reconstructs the input data from this encoded representation. During training, the network learns to capture meaningful patterns, making autoencoders well-suited for anomaly detection.

3.2 Advantages for Anomaly Detection

Autoencoders excel in capturing non-linear relationships within data, enabling them to detect complex patterns inherent in fraudulent transactions. The unsupervised nature of autoencoders allows them to learn from the intrinsic structure of the data without relying on labeled instances of fraud. Furthermore, autoencoders automatically learn relevant features from the data, enabling them to identify subtle anomalies that may not be explicitly defined.

4 Autoencoder Variants for Anomaly Detection

- **Normal Autoencoders:** Normal autoencoders, with the primary objective of minimizing the reconstruction error, prove effective in capturing normal patterns during training.
- **Sparse Autoencoders:** Sparse autoencoders introduce sparsity constraints during training, emphasizing the most critical features in the latent space. This results in improved interpretability and a focus on rare occurrences that might indicate anomalies.
- **Variational Autoencoders (VAEs):** VAEs incorporate probabilistic modeling, allowing for the generation of diverse latent representations. This enhances their ability to handle uncertainty in data, a crucial aspect for anomaly detection in dynamic fraud scenarios.
- **Denoising Autoencoders:** Denoising autoencoders, a specialized variant, are designed to reconstruct clean input data from noisy input samples. During training, random noise is introduced to the input data, forcing the autoencoder to learn robust features and filter out irrelevant noise. This not only enhances the generalization capabilities of the model but also allows it to focus on the most salient features while disregarding noisy artifacts.

5 Methodology

5.1 Dataset

The dataset includes unlabeled instances of both normal and fraudulent credit card transactions. Autoencoder Architectures: Four autoencoder architectures—VAEs, sparse autoencoders, normal autoencoders, and denoising autoencoders—are employed in the study.

Autoencoder Training (Unsupervised): Each autoencoder architecture is subjected to unsupervised training using the normal instances from the training set. The objective here is to expose the autoencoder to a comprehensive

representation of normal patterns within credit card transactions. During training, the autoencoder learns to minimize the reconstruction error, capturing the underlying structure of normal data.

Threshold Determination: The distribution of reconstruction errors for normal instances in the test set is subjected to detailed analysis.

$$Error = \sum_{i=1}^{n} (X_i - \hat{X}_i)^2$$

Statistical measures, such as the mean plus a multiple of the standard deviation, are utilized to set a threshold.

$$Threshold = Mean + Constant \times StandardDeviation$$

This threshold serves as a decisive criterion for classifying instances as normal or anomalous. Instances with reconstruction errors beyond the threshold are flagged as potential anomalies. Mean and Variance Analysis: Complementary to threshold determination, mean and variance analyses are conducted on the distribution of reconstruction errors for normal instances in the test set. This provides insights into the central tendency and variability of the reconstruction errors. Understanding the distribution characteristics aids in fine-tuning the anomaly detection process for optimal performance. Evaluation Metrics: The performance of the model is evaluated using a comprehensive set of metrics, including precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve. The threshold is fine-tuned during evaluation to achieve a balanced trade-off between false positives and

5.2 Principle of Case 1: Unsupervised Normal Data Objective:

The primary goal of Case 1 is to train an autoencoder using only normal instances of credit card transactions and then use this trained model for anomaly detection. Case 1 leverages the inherent patterns of normal credit card transactions learned during unsupervised training. By focusing exclusively on normal instances, the autoencoder becomes proficient in capturing the features of legitimate transactions. The subsequent anomaly detection process, facilitated by threshold analysis and evaluation metrics, allows the model to effectively identify transactions that deviate significantly from the learned normal patterns, signaling potential fraud or anomalies.

5.3 Principle of Case 2: Mixed Data (Two-Step Training Model) Two-Step Training Model:

In Case 2, a novel Two-Step Training Model is introduced to enhance the autoencoder's ability to detect anomalies in a mixed dataset containing both normal and anomalous instances.

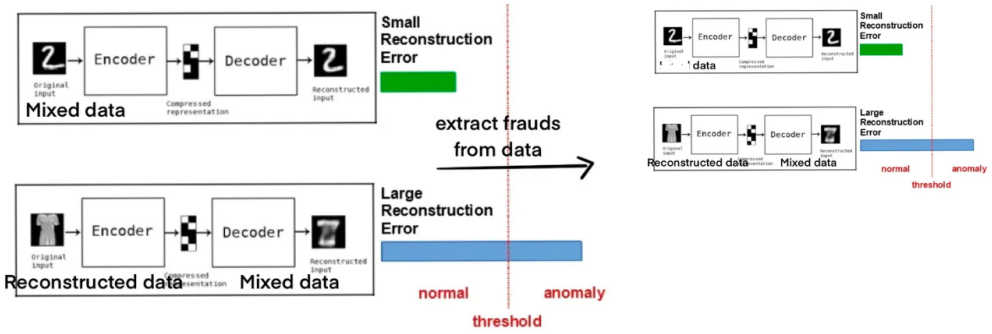

Figure 1: 2 Phase Autoencoder.

Step 1: Initial Training on Entire Dataset The first step involves training the autoencoder on the entire dataset, which includes both normal and anomalous instances. During this initial training phase, the autoencoder learns the underlying patterns present in the complete dataset. To ensure the quality of training data, data cleaning processes are applied, addressing issues such as missing values and outliers. This step aims to provide the autoencoder with a comprehensive understanding of the dataset's characteristics.

Step 2: Fine-Tuning on Normal Instances Following the initial training, the pre-trained autoencoder undergoes a fine-tuning process using only normal instances. This step is crucial for enhancing the model's ability to distinguish normal patterns from anomalies. By focusing the training exclusively on normal instances, the autoencoder refines its ability to recognize legitimate transaction patterns with greater precision.

5.4 Anomaly Detection (Mixed Data):

Once the Two-Step Training Model is completed, the autoencoder is ready for anomaly detection in the mixed dataset. This dataset contains a combination of both normal and anomalous instances.

The Two-Step Trained Autoencoder is utilized to reconstruct instances from the test set, capturing the inherent patterns learned during both the initial training on the entire dataset and the subsequent fine-tuning on normal instances. The reconstruction error for each instance is then calculated as the difference between the input and the reconstructed output. This reconstruction error serves as a crucial metric in identifying anomalies within the mixed dataset.

Results obtained from Case 2 are then compared with those from Case 1 (unsupervised normal data training) to analyze the impact of the Two-Step Training Model on anomaly detection in the mixed data scenario.

In essence, Case 2 introduces a sophisticated training approach where the autoencoder is initially exposed to the entire dataset, followed by a specialized fine-tuning process. This two-step model enhances the autoencoder's ability to discern anomalies, particularly in scenarios where mixed data patterns pose challenges to traditional anomaly detection methods.

6 Experimental Results

6.1 Evaluation Metrics

Performance metrics, including precision, recall, F1-score, and the area under the ROC curve, are employed to comprehensively evaluate the effectiveness of the autoencoder models in both Case 1 and Case 2. The evaluation metrics, including precision, recall, F1-score, and the area under the ROC curve, offer a comprehensive assessment of the autoencoder models in both Case 1 and Case 2.

6.2 Case 1 Results

Case 1 demonstrates the efficacy of autoencoders in detecting anomalies when trained solely on normal data. The results showcase the models' ability to capture and identify deviations from normal credit card transaction patterns.

Methods	Autoencoder	Sparse Autoencoder	Denoising Autoencoder	Variational Autoencoder	Ensemble	Isolation Forest
normal scaling	**0.87/0.87/0.87**	0.76/0.82/0.76	0.61/0.49/0.55	0.66/0.59/0.63		
robust scaling	0.83/0.91/0.87	0.90/0.83/0.86	0.88/0.88/0.88	0.87/0.87/0.87	**0.83/0.91/0.87**	0.3/0.41/0.35

Figure 2: Case 1 Results.

6.3 Case 2 Results

In Case 2, the Two-Step Training Model significantly improves the autoencoders' performance in the presence of mixed data. The model exhibits enhanced adaptability and accuracy in detecting anomalies, showcasing the effectiveness of the two-step training process.

Methods	Autoencoder	Sparse Autoencoder	Denoising Autoencoder	Variational Autoencoder	Ensemble	Isolation Forest
One Case	0.79/0.89/0.83	0.78/0.88/0.83	0.79/0.89/0.83	0.78/0.87/0.82	**0.82/0.81/0.82**	0.2/0.45/0.28
Two Phase				**0.80/0.87/0.83**		

Figure 3: Results Case 2.

Figure 3 illustrates the impact of the two-step training process on the generated data. In the first phase, the autoencoder trained on mixed data produces reconstructed samples, which may contain artifacts or noise due to the presence of

anomalies. However, in the second phase, the autoencoder undergoes retraining with cleaned data, resulting in improved reconstructions. The cleaned data is obtained by filtering out anomalies using anomaly detection techniques or manual inspection. These refined reconstructions demonstrate the model's ability to better capture the underlying patterns in the normal data distribution, leading to enhanced performance in anomaly detection tasks.

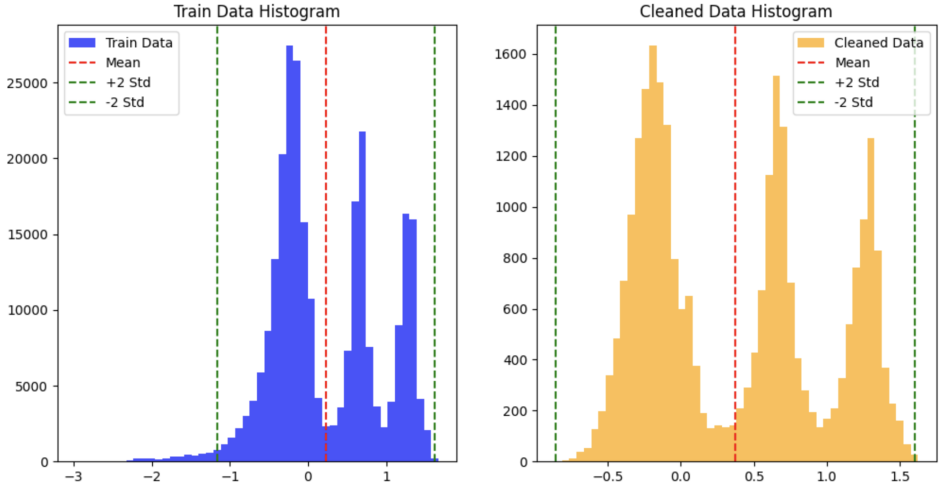

Figure 4: Mixed vs Cleaned Data after 2 Phase Step

EVALUATION

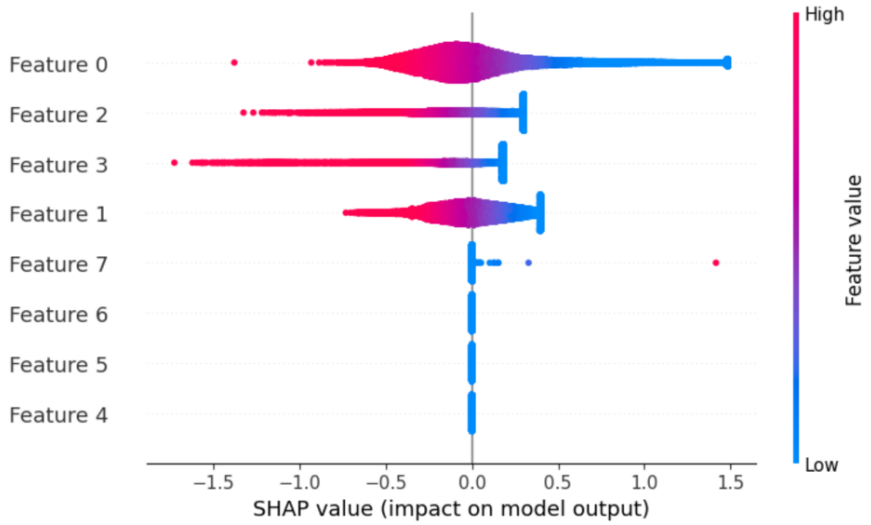

Figure 5: Evaluation.

7 Discussion of Findings

The findings underscore the importance of a nuanced approach to anomaly detection in credit card transactions. The comparative analysis, though incomplete, provides valuable insights for practitioners, aiding them in the thoughtful selection of suitable autoencoder variants tailored to the intricacies of fraud detection. Notably, the Two-Step Training Model would emerge as a powerful strategy, demonstrating its efficacy in handling mixed data scenarios. Additionally, during our analysis, we observed that only a few features exhibited significant relevance in capturing patterns indicative of fraud. This observation opens up possibilities for a more focused investigation into feature engineering, emphasizing

the identification and prioritization of these influential features. By understanding the specific attributes that contribute most to anomaly detection, practitioners can refine their models and potentially achieve even greater accuracy. Furthermore, our results highlight the adaptability and robustness of autoencoders in detecting anomalies within credit card transactions. This reinforces their pivotal role as integral tools in the ongoing battle against credit card fraud.

7.1 Future Work

The research opens avenues for future exploration, suggesting potential directions such as integrating other neural network architectures, exploring additional feature engineering techniques, and assessing the scalability of the proposed models for large-scale credit card transaction datasets. These directions aim to further enhance the applicability and effectiveness of autoencoders in real-world fraud detection scenarios.As we move forward, the exploration of hyperparameter tuning stands out as a promising avenue for improvement. Allocating dedicated efforts to fine-tune hyperparameters can potentially enhance model performance, further refining the accuracy and reliability of autoencoder-based fraud detection systems. This iterative process ensures that the models evolve to effectively adapt to the dynamic nature of fraudulent activities, contributing to the continual improvement of security measures in financial transactions.

References

[1] Charu C. Aggarwal. *Outlier Analysis*, 2nd ed. Springer, 2017.

[2] Fast Forward Labs. "An outlier is an observation generated by a different mechanism." Available online: `https://ff12.fastforwardlabs.com/#:~:text=âĂœAn%20outlier%20is%20an%20observation,generated%20by%20a%20different%20mechanism.âĂİ&text=Anomalies%2C%20often%20referred%20to%20as,a%20notion%20of%20normal%20behavior`.

[3] Guansong Pang, Chunhua Shen, Longbing Cao, Anton van den Hengel. "Deep Learning for Anomaly Detection: A Review." *Data Engineering and Software Engineering*, December 2020.

[4] Sumit Misra, Soumyadeep Thakur, Manosij Ghosh, Sanjoy Saha. "An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction." *Procedia Computer Science*, Volume 167, 2020, Pages 254-262. DOI: `10.1016/j.procs.2020.03.219`.