Fraud Detection using an Autoencoder and Variational Autoencoder

Aabisheg T (es21btech11001)
Ranveer Sahu (es21btech11025)
S Jagadeesh (es21btech11026)
Subiksha Gayathiri KK (es21btech11031)
Bhende Adarsh Suresh (cs21btech11008)

1 Abstract

Fraud detection is the process of identifying and preventing fraudulent activities within a system or organization. It plays a critical role in maintaining trust and integrity in financial transactions, healthcare, e-commerce, and various other sectors. Fraud detection is crucial in financial systems, particularly in credit card transactions, where fraudulent activities can lead to significant financial losses and damage to consumer trust. To address this challenge, we aim to develop a sophisticated model leveraging autoencoders and variational autoencoders. These neural network architectures are adept at identifying anomalies in data patterns, making them well-suited for detecting fraudulent transactions.

2 Problem Introduction

The dataset consists of credit card transactions, totaling 284,807 instances, with a small fraction of 492 transactions labeled as fraudulent. Each transaction is described by 30 attributes, including 28 principal components derived from the original features, representing the transaction details. Additionally, the dataset includes information about the time elapsed since the first transaction in the dataset and the amount paid for each transaction. With a highly imbalanced class distribution, where fraudulent transactions constitute only a small proportion of the total, the dataset poses a challenge for fraud detection algorithms, requiring robust techniques capable of accurately identifying fraudulent activities amidst the overwhelming majority of legitimate transactions.

3 Architecture of Autoencoders

Autoencoders are neural network architectures composed of an encoder and a decoder. The encoder compresses the input data into a lower-dimensional representation, called the latent space, while the decoder reconstructs the original input from this representation. Typically, autoencoders consist of multiple layers of neurons, with the middle layer representing the bottleneck or latent space. During training, the model learns to minimize the reconstruction error between the input and the output.

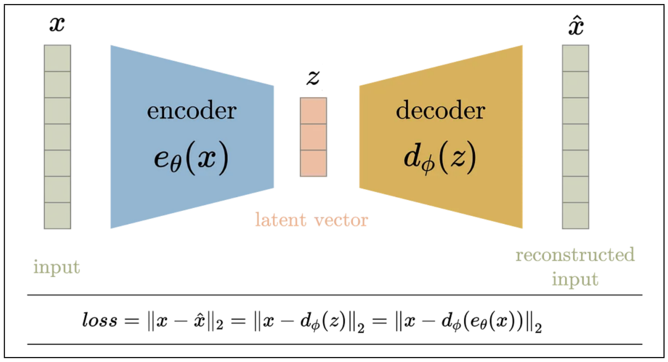

Figure 1: Architecture of Autoencoders

4 Architecture of Variational Autoencoders

Variational autoencoders (VAEs) are a variant of autoencoders that introduce probabilistic elements to the model. VAEs not only learn to encode data into a latent space but also learn the probability distribution of the latent space. This allows for generating new data points by sampling from the learned distribution. VAEs consist of two main parts: the encoder, which maps input data to a probability distribution in the latent space, and the decoder, which generates outputs from sampled latent space points.

5 Approach

Below is the step-by-step approach on how to detect anomalies using autoencoders and variational autoencoders:

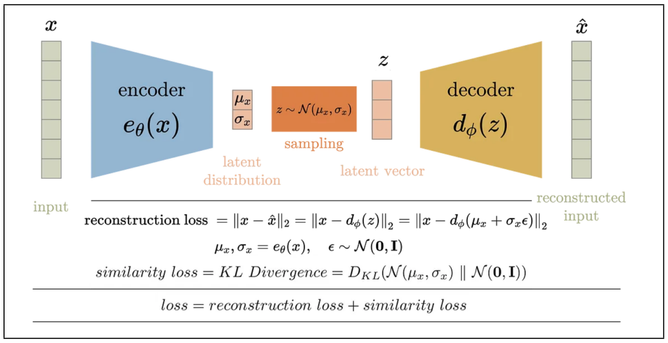
x
$\hat{x}$
z
encoder
$e_\theta(x)$
μ_x
σ_x
$z \sim \mathcal{N}(\mu_x, \sigma_x)$
sampling
decoder
$d_\phi(z)$
latent
distribution
latent vector
input
reconstructed
input
reconstruction loss $= \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(\mu_x + \sigma_x \epsilon)\|_2$
$\mu_x, \sigma_x = e_\theta(x), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$similarity\ loss = KL\ Divergence = D_{KL}(\mathcal{N}(\mu_x, \sigma_x) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}))$
$loss = reconstruction\ loss + similarity\ loss$

Figure 2: Architecture of Variational Autoencoders

- Train an autoencoder and a variational autoencoder exclusively on normal credit card transaction data to minimize reconstruction error.

- Utilize the trained models to reconstruct all credit card transactions, calculating the reconstruction error for each transaction.

- Establish a fixed threshold for the reconstruction error, beyond which transactions are deemed anomalies.

- Classify transactions surpassing the threshold as fraudulent, flagging them for further investigation.

- Compare the performance of both autoencoders and variational autoencoders in detecting fraudulent transactions based on their respective reconstruction error thresholds.

6 Evaluation Metrics:

The F1 score is a popular evaluation metric that combines precision and recall, providing a balanced measure of a model's performance. It is particularly useful in binary classification tasks, such as fraud detection, where both false positives and false negatives need to be minimized.

For both autoencoders and variational autoencoders, the F1 score is calculated based on the model's ability to correctly identify fraudulent transactions while minimizing misclassifications. Higher F1 scores indicate better balance between precision and recall, reflecting the model's effectiveness in detecting anomalies accurately. Comparing the F1 scores of autoencoders and variational autoencoders provides insights into their respective capabilities in fraud detection tasks.

7 Reconstruction loss:

In the context of autoencoders and variational autoencoders (VAEs), reconstruction loss refers to the measure of how well the model can reconstruct its input data.

7.1 Autoencoders:

In a traditional autoencoder, the reconstruction loss is typically computed as the difference between the input data and the output data produced by the decoder network. The below equation defines the loss function of the autoencoders:

$$\textbf{Loss}_{\textbf{autoencoder}} = ||\textbf{input} - \ \textbf{output}||^2$$

The goal of training an autoencoder is to minimize this reconstruction loss, encouraging the model to learn a compact representation of the input data in the latent space and effectively reconstruct the original input from that representation.

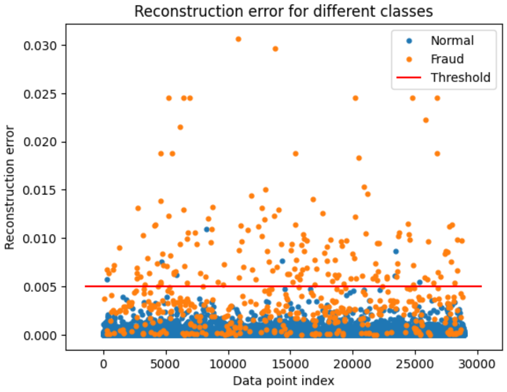

Figure 3: Reconstruction error for fraud and normal class in autoencoder. The horizontal red line denotes the threshold

7.2 Variational Autoencoders:

AEs are a type of autoencoder that incorporates probabilistic modeling into the latent space. In addition to minimizing the reconstruction loss, VAEs also aim to learn a latent space that follows a specific probability distribution, typically a Gaussian distribution. The reconstruction loss in VAEs consists of two components: the reconstruction error, which measures the difference between the input data and the output data, and the Kullback-Leibler (KL) divergence, which quantifies how closely the learned latent distribution matches the desired distribution. The overall loss function in VAEs is a combination of these two components, and the model is trained to minimize this combined loss.The below equation defines the loss function of the variational autoencoders:

$$\textbf{Loss}_{\textbf{VAE}} = \textbf{reconstruction\_loss} + \textbf{KL\_divergence}$$

As we can see in Figure 5, when training an autoencoder, the model learns to encode and decode

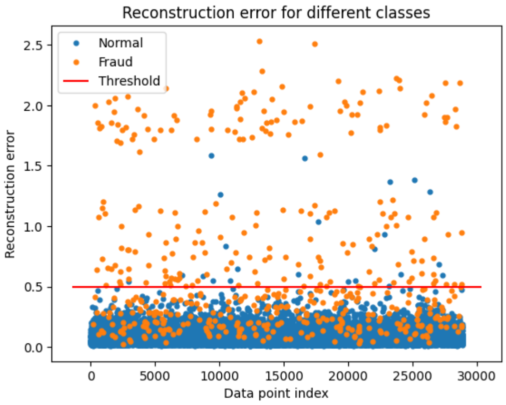

Figure 4: Reconstruction error for fraud and normal class in variational autoencoder. The horizontal red line denotes the threshold

normal data patterns effectively, resulting in low reconstruction loss for normal data samples. However, anomalies or outliers in the data may not conform to the learned patterns, causing the model to struggle to accurately reconstruct them. As a result, the reconstruction loss for data samples with anomalies tends to be higher compared to normal data samples.

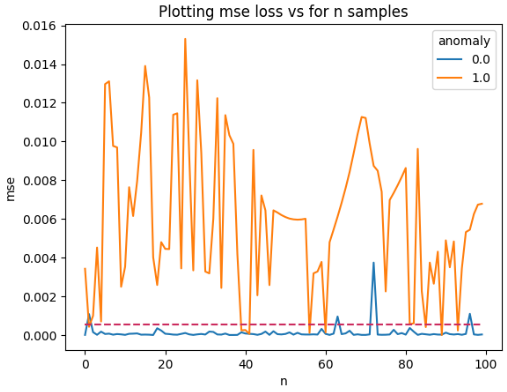

Figure 5: A graph denoting the difference in the MSE loss for data samples with anomalities and normal ones.

8 Insights from the confusion matrix:

A confusion matrix is a table that summarizes the performance of a classification model by comparing its predictions against the actual labels in a dataset. It consists of four quadrants: true positives, false positives, true negatives, and false negatives. By examining the values in these quadrants, one can gain insights into the model's accuracy, precision, recall, and overall performance in classifying instances into different categories.

8.1 Confusion matrix of autoencoder model:

The high number of true positives indicates that the model is effectively detecting fraudulent activities in the dataset. However, the presence of false positives and false negatives indicates that the model is not perfect and may misclassify some instances.

In this case, the model appears to have a relatively high false positive rate, as indicated by the 328 false positives compared to the 28,424 true positives. This means that there is a risk of incorrectly flagging some legitimate activities as fraudulent.

Additionally, the low number of true negatives suggests that the model is struggling to correctly identify non-fraudulent activities, leading to a high false negative rate. This means that some fraudulent activities may go undetected by the model.

Overall, while the model may be effective at identifying certain fraudulent activities, it is important to consider the trade-offs between false positives and false negatives and potentially adjust the model's threshold or features to improve its performance.

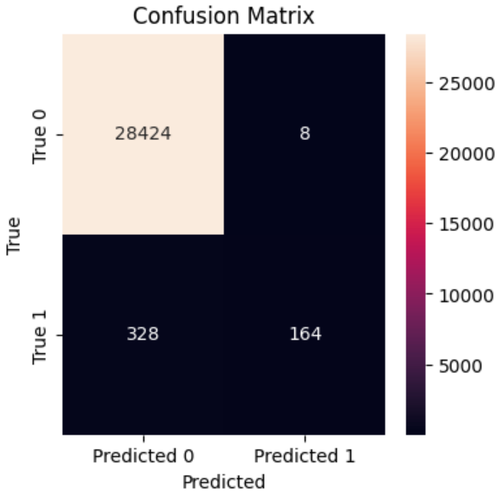

Figure 6: Confusion matrix of autoencoder model

8.2 Confusion matrix of variational autoencoder model:

The confusion matrix suggests that the variational autoencoder model correctly identified 28,409 instances of fraudulent activities (true positives) but misclassified 281 non-fraudulent instances as fraudulent (false positives). Additionally, the model accurately identified only 23 non-fraudulent instances (true negatives) and missed 211 instances of fraudulent activities (false negatives). This indicates a relatively high false positive rate and a significant number of false negatives, suggesting room for improvement in the model's performance.

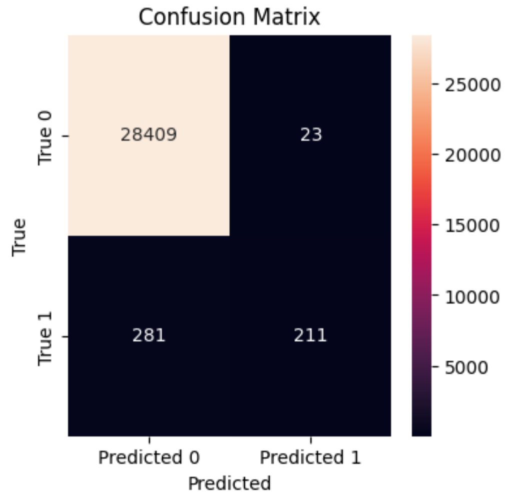

Figure 7: Confusion matrix of variational autoencoder model

Comparing the two confusion matrices, it's clear that confusion matrix 1 represents a model that is better at detecting fraudulent activities. This conclusion is based on the higher number of true positives (28,424) and lower number of false negatives (164) compared to confusion matrix 2. Additionally, confusion matrix 1 has a slightly higher true negative count (8) despite a higher false positive count (328), suggesting a more balanced performance overall.

9 Comparison of Performance using F1 Score

The F1 score of the autoencoder model is 0.4939, while the F1 score of the variational autoencoder is 0.5812. A higher F1 score indicates better overall performance in terms of both precision and recall. Since the variational autoencoder has a higher F1 score, it is considered to be better at detecting fraudulent activities compared to the autoencoder. The F1 score provides a balanced measure of a model's precision and recall, making it a useful metric for comparing the performance of different models in binary classification tasks.

10 Strategies for Improving Model Performance

To further increase the performance of the models in detecting fraudulent activities, here are a few methods that can be considered:

1. Feature Engineering: Explore and engineer additional features that may better capture the underlying patterns of fraudulent transactions. This could involve creating new features based on domain knowledge or exploring interactions between existing features.

2. Resampling Techniques: Since the dataset is highly imbalanced with only a small fraction of transactions being fraudulent, consider employing resampling techniques such as oversampling the minority class (fraudulent transactions) or undersampling the majority class (non-fraudulent transactions) to create a more balanced training set.

3. Anomaly Detection Threshold Adjustment: Adjust the threshold for classifying instances as anomalies (fraudulent) based on the model's performance metrics and business requirements. This can help balance the trade-off between false positives and false negatives and optimize the model's performance.

4. Ensemble Learning: Combine multiple models, including different variations of autoencoders and variational autoencoders, into an ensemble to leverage the strengths of each model and improve overall performance through diversity and aggregation.

5. Fine-tuning Reconstruction Loss: Experiment with different loss functions or modify the reconstruction loss term in the autoencoder or variational autoencoder to better capture the specific characteristics of fraudulent transactions and improve model performance.

6. Cross-Validation and Evaluation: Perform robust cross-validation and evaluation procedures to ensure that the models generalize well to unseen data and accurately detect fraudulent activities in real-world scenarios. Consider using techniques such as stratified k-fold cross-validation and evaluating performance metrics on separate validation and test sets.

By implementing these methods and iteratively refining the models, it is possible to further increase their performance in detecting fraudulent activities in credit card transactions.

11 Conclusion

In conclusion, improving fraud detection in credit card transactions requires a multifaceted approach that involves feature engineering, resampling techniques, model selection and tuning, anomaly detection threshold adjustment, ensemble learning, fine-tuning of reconstruction loss, and rigorous cross-validation and evaluation procedures. By implementing these strategies and iteratively refining the models, it is possible to enhance the performance of fraud detection models and better protect against fraudulent activities in credit card transactions.