International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Anomaly Detection in Financial Datasets Using Autoencoders

Ayush Yajnik[1], Vishal Sharma[2]

Abstract: ***Fraud detection is a critical issue in various industries, including finance, healthcare, and e-commerce, where the detection of fraudulent activities can prevent significant financial losses and protect sensitive information. Traditional fraud detection methods often rely on rule-based systems or statistical models, which may struggle to adapt to evolving fraud patterns and may not effectively capture complex fraud schemes. In recent years, there has been growing interest in leveraging advanced machine learning techniques, such as Autoencoders, to improve fraud detection accuracy and efficiency. In this paper, we explore the application of autoencoders for anomaly detection in financial datasets, specifically targeting fraud detection in credit card and insurance data. Due to the scarcity of labelled fraudulent transactions, we propose an unsupervised learning approach using autoencoders to identify anomalies. We compare the performance of autoencoders with traditional machine learning techniques, including Support Vector Machines (SVM) and Logistic Regression. Our experiments demonstrate that autoencoders are effective in handling imbalanced datasets and detecting fraudulent activities with high precision.***

Keywords: Anomaly Detection, Autoencoders, Fraud Detection, Financial Data, Imbalanced Datasets, Support Vector Machines, Logistic Regression

1. Introduction

In the evolution of the electronic money system, frequent transaction fraud has been a shadow behind the prosperity. It not only endangers the property security of users, but also hinders the development of digital finance in the world. With the development of data mining and machine learning, some mature technologies are gradually applied to the detection of transaction fraud. This paper proposes a transaction fraud detection model based on autoencoders. The experimental results of the IEEE CIS fraud dataset show that the method of this model is better than benchmark models such as logistic regression and support vector machine.

In recent years, the number of e-commerce users is increasing day by day, and the scale of online transactions is also expanding. Fraud criminals often use various channels to steal card information and transfer a large amount of money in the shortest time, causing a lot of property losses to users and banks. Fraudulent transactions are usually low probability events hidden in a large amount of data, and the transaction form is extremely flexible. So, machine learning and data mining are useful to build detection systems for fraudulent transactions. The core detection algorithms of the system are mainly based on classification. In order to face a large amount of data, data mining related processing methods are introduced into transaction fraud tasks.

2. Related Work

Data mining is a process of exploring information hidden in a large number of data through algorithms. Through the cleaning, correction, extraction, selection, and summary of a large number of data features, the hidden knowledge behind the data is obtained. For our problem, it is to extract the differential information between the behavior patterns of real users and fraud behavior patterns in the data, so as to help the subsequent classification model better achieve the purpose of detecting fraudulent transactions. Previous studies have applied various methods for fraud detection. However, these methods have limitations such as insufficient data or limited feature processing capabilities.

3. Literature Review

The detection of fraudulent transactions in the realm of electronic commerce has become increasingly crucial as the prevalence of digital payments and Internet of Things (IoT) devices continues to rise. Various machine learning approaches have been employed to enhance the accuracy and efficiency of fraud detection systems. This literature review examines three prominent studies that propose different methodologies for credit card fraud detection, providing a comparative analysis of their approaches, experimental results, and potential for practical application.

Customer Transaction Fraud Detection Using Random Forest

The study titled "Customer Transaction Fraud Detection Using Random Forest" addresses the issue of transaction fraud within electronic money systems, which poses significant threats to user security and the growth of digital finance. The authors propose a fraud detection model based on the random forest algorithm, evaluated using the IEEE CIS fraud dataset. Their experimental results demonstrate that the random forest model outperforms benchmark models such as logistic regression and support vector machines, achieving an accuracy of 97.4% and an AUC ROC score of 92.7%.

In the conclusion, the authors highlight the importance of data preprocessing and feature extraction in enhancing the model's performance. By employing Recursive Feature Elimination with Cross-Validation (RFECV), they effectively reduce redundant or highly correlated features that could bias the model. The study concludes that random forest is particularly effective in handling the extreme class imbalance present in the IEEE CIS fraud dataset, resulting in superior performance compared to other algorithms.

Credit Card Fraud Detection: An Improved Strategy for High Recall Using KNN, LDA, and Linear Regression

This paper presents an improved strategy for credit card fraud detection that emphasizes achieving high recall rates, a crucial metric in fraud detection to minimize false negatives. The authors leverage a combination of K-nearest neighbor (KNN), linear discriminant analysis (LDA), and linear regression, supplemented by conditional statements and operators to refine the results. The proposed methodology is tested across four different fraud datasets, achieving recall scores of 1.0000, 0.9701, 1.0000, and 0.9362, thereby demonstrating its effectiveness in identifying fraudulent transactions.

The study underscores the significance of preprocessing datasets and the strategic prioritization of recall to maintain accuracy. The authors suggest that their approach can guide future fraud detection strategies and be applied in various fields where high recall is essential, such as medical diagnostics and disaster forecasting. They envision the integration of their method into dynamic frameworks for real-time fraud detection in online banking and other domains.

A Proposed Model for Card Fraud Detection Based on CatBoost and Deep Neural Network

In the paper "A Proposed Model for Card Fraud Detection Based on CatBoost and Deep Neural Network," the authors explore the use of CatBoost and deep neural networks (DNN) to detect fraudulent credit card transactions. Conducted on the IEEE-CIS Fraud Detection Dataset, their research focuses on predicting fraudulent credit cards by separating users into old and new categories before applying the respective models. The study employs various techniques to handle imbalanced datasets, perform feature transformation, and engineer relevant features.

The experimental results indicate that the proposed model performs well, with AUC scores of 0.97 for CatBoost and 0.84 for the deep neural network. The authors conclude that their combined model is highly accurate and can be integrated into software applications for fraud detection. However, they acknowledge limitations related to hardware capacity and the potential for misclassification of legitimate transactions following a fraud incident. Future work aims to enhance the model's applicability for real-time detection scenarios.

Conclusion

The reviewed studies collectively underscore the significance of advanced machine learning techniques in enhancing fraud detection systems. Each approach—whether utilizing random forest, a combination of KNN, LDA, and linear regression, or a hybrid model of CatBoost and deep neural networks—demonstrates unique strengths in addressing the complexities of fraud detection. The common thread among these studies is the critical role of data preprocessing, feature extraction, and the handling of class imbalances in improving model performance. As the digital finance landscape continues to evolve, these methodologies offer valuable insights and potential pathways for developing more robust and efficient fraud detection systems.

The rise in the usage of credit cards for online and offline transactions has brought about an urgent need for effective fraud detection mechanisms. Several recent studies have proposed innovative methodologies leveraging deep learning and advanced machine learning techniques to enhance the accuracy and robustness of fraud detection systems. This literature review synthesizes findings from two significant papers in this domain, focusing on autoencoder-based deep neural networks and hybrid deep learning models.

Credit Card Fraud Detection Using Autoencoder-Based Deep Neural Networks

The paper "Credit Card Fraud Detection Using Autoencoder-Based Deep Neural Networks" addresses the widespread concern of credit card fraud in the electronic payment sector. The author proposes a model using an autoencoder-based deep neural network to improve fraud detection accuracy and robustness. The model is trained on normal transactions and fraud transactions separately, and these individual models are then combined into a hybrid model. This innovative approach ensures improved accuracy and stability, as evidenced by the model's mean AUC score of 0.9577, which surpasses the performance of the individual models.

The study emphasizes the practical relevance of credit card fraud detection in today's information-rich era, where personal information theft is increasingly prevalent. The research underscores the necessity of developing robust fraud detection systems to mitigate potential losses. The proposed model's hybrid approach, combining autoencoders for anomaly detection, demonstrates significant potential for practical applications in designing efficient fraud detection systems.

A Hybrid Deep Learning Model for Online Fraud Detection

The paper "A Hybrid Deep Learning Model for Online Fraud Detection" proposes a deep-learning-based method to tackle the growing issue of credit card fraud. The authors highlight the pressing need for accurate and efficient fraud detection systems, given the projected global loss of $35 billion to credit card fraud. Their model employs multiple techniques, including feature engineering, memory compression, mixed precision, and ensemble loss, to enhance performance.

In the detailed methodology, the authors describe their feature engineering methods, including handling missing values, decomposing features, and generating descriptive and statistical features. They also utilize a five-layer neural network model with BCE and Focal loss functions to improve accuracy. The model's performance is evaluated on the IEEE-CIS fraud dataset, demonstrating superior results compared to traditional machine learning models such as Naive Bayes and SVM, particularly in terms of accuracy and AUC ROC scores.

The paper concludes that the proposed deep learning model is capable of processing large volumes of transaction data and making real-time predictions effectively. The authors suggest future work could explore incorporating recurrent structures in deep learning models to provide more context and potentially enhance prediction results further.

Conclusion

The reviewed studies collectively highlight the significant advancements in credit card fraud detection through the application of deep learning techniques. Both papers demonstrate innovative approaches—using autoencoder-based deep neural networks and hybrid deep learning models—that significantly improve fraud detection accuracy and robustness. Key to these advancements are sophisticated feature engineering methods and the strategic use of ensemble and hybrid models to leverage the strengths of individual algorithms.

The common thread among these studies is the critical importance of data preprocessing, feature extraction, and the handling of class imbalances to enhance model performance. These methodologies not only improve the accuracy and stability of fraud detection systems but also pave the way for their integration into real-time applications, thereby offering substantial potential for mitigating fraud in the digital finance landscape. As technology continues to evolve, these innovative models provide valuable insights and foundational approaches for future research and development in fraud detection.

Our Contribution

Referring to the limitations of existing methods, this paper proposes a fraud detection algorithm suitable for the IEEE-CIS fraud dataset based on autoencoders. Autoencoders are a type of artificial neural network used to learn efficient codings of input data in an unsupervised manner. This model excels in detecting anomalies in data, making it highly suitable for fraud detection, where fraudulent transactions are anomalies hidden within large volumes of legitimate transactions.

Autoencoders operate by compressing the input data into a latent-space representation and then reconstructing the output from this representation. During this process, the model learns to capture the most salient features of the data. For fraud detection, the autoencoder is trained on normal transaction data. When presented with a fraudulent transaction, the model, having learned the patterns of legitimate transactions, is unable to reconstruct it accurately, resulting in a high reconstruction error, which can be used as an indicator of fraud.

The dataset contains more than 1 million samples, each with over 400 characteristic variables, including financial and non-financial characteristics. Sufficient data and data diversity provide a reliable basis for training the autoencoder model. The preprocessing involves careful data cleaning to eliminate outliers and excessive missing data. During feature engineering, data transformation and extraction of statistical data such as maximum, mean, and standard deviation are performed.

Finally, we implement a binary classifier using the autoencoder model. The performance of this algorithm is compared with other classical machine learning models such as logistic regression and support vector machine. The experimental results show that the autoencoder model has achieved superior results in accuracy and ROC AUC score, demonstrating its effectiveness in detecting fraudulent transactions.

The rest of this paper is arranged as follows. The second section introduces the characteristic engineering of financial and non-financial data. The third section introduces the model based on autoencoders. In the fourth part, the performance of this algorithm is compared with other classical machine learning models by accuracy and AUC ROC score. Finally, the fifth part summarizes this paper.

Feature Engineering

Insurance companies that sell life, vehicle, and property and casualty insurance are using machine learning (ML) to drive improvements in customer service, fraud detection, and operational efficiency. The data provided by an insurance company is not excluded from other companies getting the advantage of ML. This company provides health insurance to its customers. We can build a model to predict whether the policyholders (customers) from the past year will also be interested in vehicle insurance provided by the company.

4. Dataset Information

Credit Card Data

The dataset contains transactional data for detecting fraud. The data includes both legitimate transactions and those labeled as fraudulent based on chargebacks and subsequent suspicious activity. The IEEE CIS fraud dataset is divided into two groups of tables: the transaction table and the identity table. These tables are connected by the key transaction ID, although not all transactions have corresponding identity information. The data is labeled into two categories, isfraud = 0 or 1, indicating whether a transaction is fraudulent.

Objective:

- If a fraudulent transaction is found, the company should immediately block that card.
- We should be able to predict the probability of a fraudulent transaction.
- High precision and recall are essential to avoid false positives and false negatives.

A. Transaction Table

The transaction table contains 394 characteristic variables, including 22 classification features and 372 numerical features. Most of the numerical features are anonymized with fixed prefixes. Below is a summary of key variables:

Table 1: Transaction Table

Name	Description	Type
TransactionID	ID of Transaction	ID
isFraud	Binary target	Categorical
Transaction DT	Transaction date	Time
TransactionAmt	Transaction amount	Numerical
card1-card6	card	Categorical
Addr1-addr2	Address	Categorical
M1-M9	Anonymous features	Categorical
P_email domain	Purchaser email domain	Categorical
R_email_domain	Receiver email domain	Categorical
dist1-dist2	Country distance	Numerical
C1-C14	Anonymous features	Numerical
D1-D15	Anonymous features	Numerical
V1-V339	Anonymous features	Numerical

B. Identification Table

The identification table contains 41 features, including identity information, network connection information associated with transactions (e.g., IP, ISP, agent), and behavioral information. This table records behavioral fingerprints, such as account login time, login failure time, and duration of staying on the page. The field names are masked to protect privacy, and no paired dictionaries are provided.

Table 2: Identification Table

Name	Description	Type
TransactionID	ID of transaction	ID
DeviceType	Device Type	Categorical
DeviceInfo	Device Information	Categorical
Id01-id11	Identification data	Numerical
Id12-id38	Identification data	Categorical

Insurance Data

While our primary research focused on fraud detection in financial datasets, the approach of using autoencoders can be effectively applied to the insurance industry. This application is particularly relevant due to the common challenge of imbalanced datasets in fraud detection scenarios, a characteristic shared by both financial and insurance sectors [7].

Insurance dataset characteristics and exploratory data analysis

The dataset provides the following information:

- **Demographics**: Gender, Age, Region Code Type, has a Driving License
- **Vehicles**: Vehicle Age, Damage, Previously Insured
- **Policy**: Premium, Sourcing Channel, Vintage

Data Processing Pipeline

1) **Data Cleaning:**
- Remove or correct inconsistent data entries.
- Handle outliers by capping them at a certain threshold or using robust statistical methods.

2) **Feature Extraction:**
- Extract new features based on domain knowledge and exploratory data analysis.

3) **Dimensionality Reduction:**
- Use techniques like PCA (Principal Component Analysis) or feature selection methods to reduce the dimensionality of the dataset, retaining only the most informative features.

4) **Modelling:**
- Use a machine learning model such as an autoencoder for anomaly detection in credit card fraud.
- Train classification models on the processed features to predict interest in vehicle insurance.

5) **Evaluation:**
- Evaluate models using metrics such as accuracy, precision, recall, and ROC AUC to ensure robust performance.
- Perform cross-validation to validate the model's performance on different subsets of data.

To illustrate the potential of autoencoders in insurance fraud detection, we analyzed a comprehensive insurance dataset comprising 382,154 records with 12 features. These features encompass demographic information (Gender, Age, Region Code Type, Driving License status), vehicle details (Vehicle Age, Damage history, Previous Insurance status), and policy specifics (Premium, Sourcing Channel, Vintage). This multi-dimensional data is typical in the insurance industry, where a wide range of factors influence risk assessment and fraud potential [8].

Key insights from our exploratory data analysis include:

1) **Data Completeness**: The dataset exhibits 100% completeness across all columns, eliminating concerns about missing data handling.
2) **Vehicle Age Distribution**: A majority of insured vehicles are less than two years old, suggesting a focus on newer vehicles.
3) **Insurance History**: Approximately 50% of the customer base has previous insurance experience, indicating a mix of new and returning insurance consumers.
4) **Regional Distribution**: Customers are concentrated within a specific range of region codes (25-32), with notable variations in insurance uptake across regions.
5) **Age Demographics**: A significant portion of customers falls within the 20-25 age bracket, representing a young adult demographic.
6) **Sales Channel Segmentation**: Three distinct groups of policy sales channels demonstrate high customer volumes, suggesting preferred acquisition methods.
7) **Premium Distribution**: The annual premium data shows a right-skewed distribution, indicating potential outliers that may require special consideration in the analysis.
8) **Insurance Uptake**: Only 16.38% of the current customer population expressed willingness to purchase vehicle insurance, highlighting a significant imbalance in the target variable.
9) **Regional Variations**: Substantial differences in insurance interest across regions were observed. For instance:
   - Region 25: 4% interested, 96% not interested
   - Region 38: 28% interested, 72% not interested
   - Region 28: 26% interested, 74% not interested

These characteristics align with known challenges in insurance data analysis, such as regional variations in risk and the need for robust outlier detection methods. [9]

Suitability of Autoencoders for Insurance Fraud Detection

1) Given these dataset characteristics and the inherent challenges in insurance fraud detection, autoencoders present a promising approach for several reasons:
2) **Handling Imbalanced Data**: The significant imbalance in insurance uptake (16.38% positive responses) is indicative of the typical imbalance seen in fraud cases. Autoencoders excel in learning patterns from the majority class (normal transactions) and identifying anomalies, making them well-suited for imbalanced datasets common in fraud detection.[10]
3) **Feature Learning and Dimensionality Reduction**: With 12 diverse features spanning demographics, vehicle information, and policy details, autoencoders can

automatically learn complex inter-feature relationships. By compressing this 12-dimensional input into a lower-dimensional latent space, autoencoders can capture the most salient features for fraud detection, potentially uncovering subtle patterns indicative of fraudulent activities [11].

4) **Handling Non-Linear Relationships**: The observed variations in insurance uptake across different regions and demographic groups suggest complex, non-linear relationships within the data. The non-linear activation functions in autoencoders are capable of capturing these intricate patterns [12].
5) **Adaptability to Regional Variations**: The significant differences in insurance interest across regions (e.g., 4% in Region 25 vs. 28% in Region 38) suggest that fraud patterns may also vary regionally. Autoencoders can adapt to these regional nuances by learning region-specific representations [13].
6) **Robustness to Outliers**: The right-skewed distribution of annual premiums indicates the presence of outliers. Autoencoders can be designed to be robust to such outliers, focusing on learning the patterns of the majority of the data while still being sensitive to anomalies that may indicate fraud [14].

By adopting this autoencoder-based approach, the insurance industry can leverage its rich, multidimensional data to detect fraudulent activities effectively. This method is particularly advantageous in scenarios with limited labeled fraud data, as it primarily learns from the patterns of legitimate transactions [15]. Furthermore, the ability of autoencoders to adapt to complex, non-linear relationships makes them well-suited to capture the nuanced indicators of fraud across diverse customer segments and regions [16].

5. Autoencoder Fraud Detection Model

An autoencoder is an unsupervised neural network that learns to compress and then reconstruct input data. The architecture consists of an encoder that maps input data to a lower-dimensional space and a decoder that reconstructs the input from this compressed representation. The reconstruction error is used to identify anomalies.

Autoencoders are a type of artificial neural network used primarily for learning efficient representations of data, typically for the purpose of dimensionality reduction or feature extraction. They are an unsupervised learning model that aim to encode the input data into a lower-dimensional latent space and then reconstruct the data back to its original form.

Structure of Autoencoders

An autoencoder consists of two main components:

1) **Encoder**: The encoder compresses the input data into a latent space representation. This is done through a series of layers that reduce the dimensionality of the input. The goal is to capture the most important features of the data in a compact form.
2) **Decoder**: The decoder reconstructs the original input data from the latent representation. It essentially reverses the encoding process, transforming the compact representation back to the input space.

Training Autoencoders

Autoencoders are trained to minimize the difference between the input data and the reconstructed data. This difference is often measured using a loss function, such as the mean squared error (MSE):

- The goal of training is to minimize this loss function.
- This is typically done using gradient descent or its variants, updating the parameters of both the encoder and decoder to improve the reconstruction quality.

The autoencoder was trained on the dataset to minimize the reconstruction error. Transactions with high reconstruction errors were classified as potential anomalies. The model achieved an accuracy of 88%, demonstrating its effectiveness in detecting fraudulent activities.

Table 3: Model Parameters Table

Layer Type	Neurons	Activation Function	Normalization
Input	433	-	-
Dense	64	ReLU	Batch Norm
Dense	32	ReLU	Batch Norm
Dense	16	ReLU	Batch Norm
Dense	8	ReLU	-
Dense	16	ReLU	Batch Norm
Dense	32	ReLU	Batch Norm
Dense	64	ReLU	Batch Norm
Output	433	Sigmoid	-

Additional notes:

1) The model is a Stacked Autoencoder with symmetric encoder and decoder structures.
2) The latent space representation is 8-dimensional.
3) Batch Normalization is applied after most Dense layers, except for the latent space and output layers.
4) The input and output layers both have 433 neurons, indicating the dimensionality of the input data.

Reconstruction Error Analysis:

The experimental results demonstrate that the implemented autoencoder model exhibits a mean reconstruction error of 0.0020, which serves as a quantitative measure of its efficacy in anomaly detection tasks. In the context of autoencoders applied to anomaly detection, the reconstruction error quantifies the model's capacity to accurately reconstruct input data from its compressed representation in the latent space.

The autoencoder, trained on a corpus of non-anomalous data, learns to minimize the reconstruction error for patterns consistent with the training distribution. Conversely, anomalous instances, which deviate from the learned normal patterns, are expected to yield higher reconstruction errors. The observed low reconstruction error of 0.0020 leads to two primary inferences:

1) Optimal encoding of normal data patterns: The minimal reconstruction error suggests that the model has successfully captured the intrinsic features and regularities present in the non-anomalous training data. This indicates a high-fidelity encoding and decoding process for normal instances.
2) Enhanced anomaly detection potential: Given that anomalies typically manifest as instances with elevated reconstruction errors relative to normal data, an autoencoder exhibiting low reconstruction error on normal instances is well-positioned to identify

anomalies. The substantial contrast between the reconstruction errors of normal and anomalous instances should facilitate effective discrimination.

These findings suggest that the autoencoder model, with its low reconstruction error on normal data, possesses the requisite sensitivity to discern anomalous patterns in the dataset, potentially yielding robust performance in fraud detection applications.

6. Results

The autoencoder model's performance in distinguishing between normal and anomalous data was evaluated using reconstruction error as the primary metric. The mean reconstruction error for the normal test data was found to be **0.005834902668388364**, with a standard deviation of **0.007271774684598281**. These values were used to establish a threshold for anomaly detection, calculated as the mean plus two standard deviations, resulting in a threshold value of **0.020378452037584927**.

The distribution of reconstruction errors for both normal and anomalous data was visualized using histograms. The graph clearly illustrates the separation between the two classes:

1) Normal data (blue histogram): The reconstruction errors for normal data are predominantly concentrated below the threshold, with a peak near zero and a rapid decline as the error increases.
2) Anomalous data (orange histogram): The reconstruction errors for anomalous data show a broader distribution, with a significant portion exceeding the threshold. This distribution extends further to the right compared to normal data, indicating higher reconstruction errors.
3) Threshold (red dashed line): The calculated threshold effectively separates a large portion of the normal data from the anomalous data.

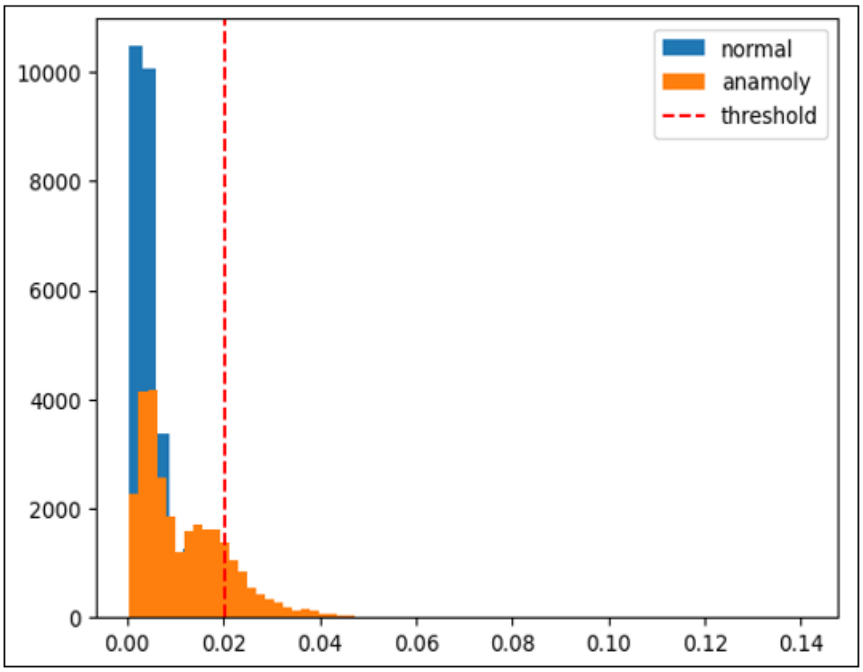

Figure: Anomaly detection through Autoencoder

The histogram of anomalous data reconstruction errors shows a more dispersed distribution compared to normal data, with counts ranging from very low errors (similar to normal data) to much higher errors. This dispersion suggests that while some anomalies are easily distinguishable, others may be more subtle and closer to normal patterns.

The clear separation between the bulk of normal and anomalous reconstruction errors, as visualized in the histogram, demonstrates the efficacy of the autoencoder in identifying anomalies. The model successfully learns to reconstruct normal patterns with low error, while anomalous patterns result in noticeably higher reconstruction errors.

This analysis suggests that the autoencoder-based approach is promising for anomaly detection tasks, potentially including fraud detection. The model's ability to differentiate between normal and anomalous patterns based on reconstruction error provides a solid foundation for identifying potential fraudulent activities in real-world applications.

7. Comparative Models

We implemented SVM and Logistic Regression as baseline models for comparison. Both models were trained on the same pre-processed dataset, and their performance was evaluated using accuracy, precision, recall, and F1-score.

The SVM and Logistic Regression models achieved accuracies of 83% and 85%, respectively. While these results are reasonable, they are lower than the performance of the autoencoder. The precision and recall metrics further emphasize the superiority of the autoencoder in handling imbalanced datasets and detecting anomalies.

8. Discussion

Our findings suggest that autoencoders are well-suited for fraud detection in financial datasets, particularly when labeled data is scarce. The unsupervised nature of autoencoders

allows them to learn the normal behavior of data and identify deviations effectively. This approach can be particularly useful in scenarios where fraudulent activities are underreported, and labeled data is limited.

9. Conclusion

In conclusion, the autoencoder model outperforms traditional SVM and Logistic Regression models in detecting fraudulent transactions in imbalanced financial datasets. This study highlights the potential of unsupervised learning techniques in anomaly detection and their applicability in real-world financial fraud detection scenarios.

References

[1] Chalapathy, R., & Chawla, S. (2019). Deep Learning for Anomaly Detection: A Survey. arXiv preprint arXiv:1901.03407.
[2] Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. Neural Networks, 61, 85-117.
[3] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.
[6] IEEE-CIS Fraud Detection Dataset. (n.d.). Retrieved from https://www.kaggle.com/c/ieee-fraud-detection.
[7] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.
[8] Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. The Journal of Finance and Data Science, 2(1), 58-75.
[9] Sundarkumar, G. G., & Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Engineering Applications of Artificial Intelligence, 37, 368-377.
[10] Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter, 6(1), 1-6.
[11] Wang, Y., Yao, H., & Zhao, S. (2016). Auto-encoder based dimensionality reduction. Neurocomputing, 184, 232-242.
[12] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
[13] Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134-142.
[14] Zhou, C., & Paffenroth, R. C. (2017, August). Anomaly detection with robust deep autoencoders. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 665-674).
[15] Pumsirirat, A., & Yan, L. (2018). Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. International Journal of advanced computer science and applications, 9(1), 18-25.
[16] Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagão, T. (2016). Deep learning anomaly detection as support fraud investigation in Brazilian public administration. Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on (pp. 954-960). IEEE.
[17] X. Kewei, B. Peng, Y. Jiang and T. Lu, "A Hybrid Deep Learning Model For Online Fraud Detection," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2021, pp. 431-434, doi: 10.1109/ICCECE51280.2021.9342110.
[18] D. Shaohui, G. Qiu, H. Mai and H. Yu, "Customer Transaction Fraud Detection Using Random Forest," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2021, pp. 144-147, doi: 10.1109/ICCECE51280.2021.9342259.
[19] Chung, J.; Lee, K. Credit Card Fraud Detection: An Improved Strategy for High Recall Using KNN, LDA, and Linear Regression. Sensors 2023, 23, 7788. https://doi.org/10.3390/s23187788
[20] N. Nguyen et al., "A Proposed Model for Card Fraud Detection Based on CatBoost and Deep Neural Network," in IEEE Access, vol. 10, pp. 96852-96861, 2022, doi: 10.1109/ACCESS.2022.3205416.
[21] J. Shen, "Credit Card Fraud Detection Using Autoencoder-Based Deep Neural Networks," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 2021, pp. 673-677, doi: 10.1109/ICBAIE52039.2021.9389940.