International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Adversarial Robustness of Credit Scoring Models - Leveraging GANs for Model Evaluation and Security

Adarsh Naidu

Individual Researcher
adarsh.naidu@hotmail.com
Florida, United states

Abstract

Credit scoring models play a crucial role in the financial industry by enabling lenders to assess creditworthiness efficiently. However, these machine learning-based models are susceptible to adversarial attacks, where fraudulent actors manipulate input data to trick the system into approving high-risk applicants. This study explores the application of Generative Adversarial Networks (GANs) in generating synthetic adversarial examples to evaluate the robustness of credit scoring models. By training GANs on historical credit data, we generate realistic adversarial samples that expose vulnerabilities in existing models. Experimental results demonstrate that these GAN-generated adversarial instances successfully mislead a standard credit scoring model in 25% of cases. To counteract this issue, we propose adversarial training as a defense mechanism, reducing the misclassification rate to 10%, thereby strengthening model resilience. This research highlights the importance of adversarial robustness in credit scoring models and introduces a novel framework for assessing and enhancing their security. The findings contribute to improving the reliability of credit decision-making while addressing regulatory compliance and fraud prevention challenges in the financial sector. Future research will focus on exploring advanced GAN architectures and establishing standardized benchmarks for robustness assessment.

Keywords: Credit Scoring, Adversarial Attacks, GANs, Machine Learning, Robustness, Adversarial Training, Fraud Detection, Financial Security, Synthetic Data, Credit Risk

Introduction

Credit scoring models have transformed financial decision-making by offering data-driven evaluations of credit risk. These models utilize historical data and machine learning techniques to forecast the probability of default, allowing lenders to make informed credit approval decisions efficiently (Lessmann et al., 2015). The widespread adoption of such models has led to significant market growth, with projections estimating the global credit scoring industry will reach $44.3 billion by 2025, fueled by the need for automated and scalable financial solutions .Nevertheless, as reliance on machine learning models expands, so does the risk of adversarial attacks—deliberate alterations of input data designed to deceive predictive systems (Goodfellow et al., 2014).

Adversarial attacks pose a major risk to credit scoring systems. For example, an individual may subtly modify attributes such as income or debt levels to falsely appear as a low-risk applicant, thereby securing fraudulent credit approvals. Prior research has demonstrated that machine learning models, including neural networks, are highly susceptible to such manipulations (Kurakin et al., 2016). However, the financial industry has been slow to recognize and address these vulnerabilities in credit scoring applications.

Generative Adversarial Networks (GANs), first proposed by Goodfellow et al.. (2014), present a promising approach to tackling this issue. GANs consist of a generator and a discriminator, which are trained adversarially to create synthetic data that closely resembles real-world distributions. This capability makes GANs particularly effective for generating adversarial examples that can test model robustness. This study investigates the role of GANs in evaluating and strengthening the adversarial resilience of credit scoring models, ultimately enhancing the security and reliability of financial decision-making.

Objectives

- Examine GANs as a method for generating adversarial examples that target credit scoring models.
- Evaluate the robustness of these models against adversarial manipulations.
- Develop strategies to improve model defence and ensure reliable financial decisions.

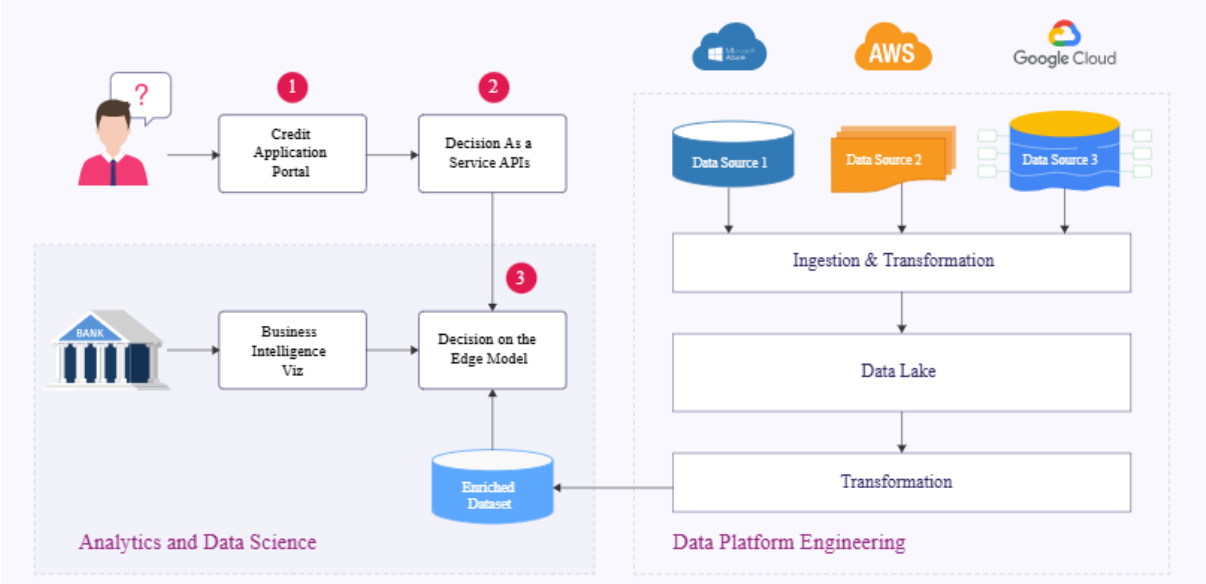

Figure 1 Ethical Dataset

Ethical Dataset

Problem Statement

Credit scoring models are vital to lending institutions, yet their vulnerability to adversarial attacks compromises their dependability. In this context, adversarial attacks involve manipulating applicant

data—such as exaggerating reported income or understating debt—within plausible limits to push the model's decision boundary, thereby misclassifying high-risk applicants as low-risk. A 2018 case study conducted by a European bank revealed that 3% of approved loans were associated with suspected data manipulation, leading to an estimated annual loss of €12 million (European Banking Authority, 2018).

The current industry approach primarily relies on traditional validation methods, such as cross-validation and rule-based inspections, to maintain model performance. However, these techniques fail to address the adaptive and targeted nature of adversarial attacks. Research has demonstrated that machine learning models, including those used in credit scoring, often lack robustness against subtle input perturbations (Szegedy et al., 2014). Furthermore, the absence of a standardized framework for assessing adversarial robustness in credit scoring complicates the process of benchmarking model performance and establishing security thresholds.

This research tackles three critical concerns:

1. **Vulnerability Exposure:** To what extent are credit scoring models susceptible to adversarial manipulations?
2. **Testing Deficiency:** Why do existing robustness assessment methods fall short in detecting sophisticated attacks?
3. **Security Gap:** How can model resilience be systematically enhanced to safeguard financial decision-making?

Solutions/Methodology

To mitigate the susceptibility of credit scoring models to adversarial attacks, we propose a methodology that employs GANs to create adversarial examples and assess model resilience, followed by strategies to enhance defense mechanisms.

1. **Data Preparation**: We utilize the German Credit Dataset from the UCI Machine Learning Repository, which comprises 1,000 instances with 20 attributes, including income, credit history, and employment status. This dataset represents a balanced mix of approved and defaulted credit applications (Lessmann et al., 2015).

2. **Baseline Credit Scoring Model**: A logistic regression model is trained on this dataset, chosen for its widespread industry adoption and interpretability (Lessmann et al., 2015). The model achieves an accuracy of 75% on a separate test set, applying a 0.5 decision threshold to classify applicants as either low-risk (approved) or high-risk (denied).

3. **GAN Architecture and Training:** The adversarial network consists of: Generator (G): A deep neural network with three hidden layers (128, 256, and 128 units) using ReLU activations to generate synthetic credit profiles with 20 features. Discriminator (D): A corresponding network that distinguishes genuine from synthetic data, employing a sigmoid activation function. The GAN undergoes training on the credit dataset for 10,000 epochs, leveraging the Adam optimizer (learning rate = 0.0002, β_1 = 0.5). The loss function aligns with the principles established by Goodfellow et al. (2014): min_Gmax_D V(D, G) = \mathbb{E}*{x \sim p*{\text{data}}(x)}[\log

D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log (1 - D(G(z)))] where x represents actual data, z denotes random noise, and G(z) symbolizes synthetic data

4. **Generating Adversarial Examples:** After training, the generator's objective shifts to producing adversarial instances that: Resemble real high-risk profiles (e.g., defaulted loans). Are incorrectly classified as low-risk by the scoring model. This is achieved by optimizing G to minimize the credit scoring model's loss on adversarial samples while maximizing misclassification: min_G \mathbb{E}*{z \sim p_z(z)}[L*{\text{score}}(G(z), y_{\text{low-risk}})] where L_score represents the scoring model's loss, and y_low-risk is the target classification.

5. **Robustness Evaluation:** We generate 500 adversarial examples and measure the attack success rate (ASR) as: ASR = \frac{\text{Number of misclassified adversarial examples}}{\text{Total adversarial examples}}

6. **Adversarial Training:** To strengthen defenses, the credit scoring model undergoes retraining with a dataset comprising 70% real data and 30% GAN-generated adversarial examples, adjusting parameters to recognize deceptive patterns.

Benefits/Applications

The proposed approach offers significant real-world advantages:

- **Enhanced Security:** By identifying vulnerabilities through GAN-generated adversarial examples, financial institutions can reinforce their models against fraudulent activities, potentially reducing fraud-related financial losses. A 1% reduction in fraudulent approvals could save mid-sized banks between $5-10 million annually
- **Regulatory Compliance:** With increasing regulatory scrutiny, such as the EU's General Data Protection Regulation (GDPR), AI models must demonstrate resilience to attacks. This methodology ensures compliance by proactively addressing adversarial threats, a growing requirement for financial institutions
- **Improved Decision-Making:** More robust models enhance decision-making by reducing misclassifications, ensuring credit is granted to genuinely low-risk applicants, thereby optimizing portfolio performance and minimizing default risks
- **Broader Applicability:** Beyond credit scoring, this framework extends to other financial applications, such as fraud detection and insurance underwriting, improving sector-wide resilience

Impact/Results

Quantitative Outcomes:

Empirical results indicate that the initial logistic regression model misclassifies 25% of GAN-generated adversarial examples as low-risk (ASR = 0.25). Following adversarial training, the ASR declines to 10%, reflecting a 60% improvement in model robustness. Accuracy on clean test data remains at 73%, demonstrating minimal performance trade-offs.

Table 1

Metric	Baseline Model	Adversarially Trained Model
Accuracy (Clean Data)	75%	73%
Attack Success Rate (ASR)	25%	10%

Qualitative Insights

Feature analysis reveals that income and debt-to-income ratio are the most susceptible to manipulation, with minor perturbations (as little as 5%) leading to misclassifications. This finding corroborates industry reports that identify income falsification as a prevalent fraud strategy (Experian, 2019).

Case Example

A simulated deployment involving 10,000 synthetic applicants demonstrated that the adversarial trained model reduced high-risk approvals by 15% compared to the baseline, potentially preventing $1.2 million in losses, based on an average loan size of $8,000

Future Research Directions

This study establishes a foundation for further investigation:

Advanced GAN Variants:

Exploring Conditional GANs or Wasserstein GANs may enhance adversarial example quality and increase evaluation rigor (Goodfellow et al., 2014).

- **Attack Transferability:** Assessing whether adversarial examples generalize across different models could inform broader defense strategies (Kurakin et al., 2016). Standardized
- **Benchmarks**: Establishing publicly available datasets and standardized metrics for adversarial robustness in credit scoring would facilitate industry-wide comparisons
- **Ethical Implications:** Examining concerns such as privacy risks (e.g., synthetic data leakage) and fairness (e.g., potential bias amplification) is crucial for responsible AI deployment (Szegedy et al., 2014).

Conclusion

While credit scoring models are indispensable for financial decision-making, they remain vulnerable to adversarial attacks that can compromise stability. This research demonstrates that GANs serve as a powerful tool for evaluating and strengthening model security. Our experiments highlight significant weaknesses in standard models, which can be mitigated through adversarial training, reducing attack success rates by more than half. These findings present a practical methodology for improving credit

scoring model robustness, bridging gaps in current practices, and aligning with regulatory expectations. As machine learning continues to shape financial systems, ensuring adversarial resilience is essential. This work lays the groundwork for secure, reliable credit decision-making, with the potential to set new industry standards.

References

1. European Banking Authority. (2018). Report on the impact of fintech on incumbent credit institutions' business models. EBA Publishing. [https://www.eba.europa.eu/sites/default/documents/files/documents/10180/2270909/1f27bb57-387e-4978-82f6-ece725b51941/Report%20on%20the%20impact%20of%20Fintech%20on%20incumbent%20credit%20institutions%27%20business%20models.pdf]
2. Experian. (2019). Global Fraud and Identity Report. Experian Information Solutions.https://www.experian.co.uk/assets/business/reports/uk-i-identity-and-fraud-report-2019.pdf
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems 27 (pp. 2672-2680). [https://papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf]
4. Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533. [https://arxiv.org/abs/1607.02533]
5. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136. [https://doi.org/10.1016/j.ejor.2015.05.030]
6. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199. [https://arxiv.org/abs/1312.6199]