Introduction and Research Question

Financial fraud detection is a serious problem in our modern economies due to the volume of online transactions and the fact thathey will only continue to increase. Electronic payment systems also contribute to this. As transaction volumes increase, so does the opportunity for fraud to take place, so it is imperative to find methods to prevent and detect these behaviors. A key challenge in fraud detection is extreme class imbalance, where fraudulent transactions represent only a very small fraction of all activity. This imbalance complicates traditional machine learning approaches and makes naïve performance metrics misleading. Many fraud detection systems must identify a handful of malicious cases among hundreds of thousands of legitimate transactions. Missing fraudulent activity can result in substantial financial losses, while overly aggressive detection can disrupt normal customers. Beyond the immediate monetary loss, the persistence of undetected fraud is one of the key factors in the erosion of trust in any digital banking infrastructure. As those perpetuating this fraud are becoming more sophisticated with the increases in technology, “rule-based systems” are starting to become increasingly more out of touch with reality. This now requires programmers and security professionals to move toward dynamic, data-driven mining techniques that can adapt to evolving patterns. As a result, designing effective fraud detection systems requires careful trade-offs. This project focuses on comparing supervised fraud detection models with unsupervised anomaly detection techniques under severe class imbalance. We will look at how these approaches fluctuate in their ability to detect fraud transactions. Understanding these differences is required for building real-world fraud detection that is both practical and scalable.

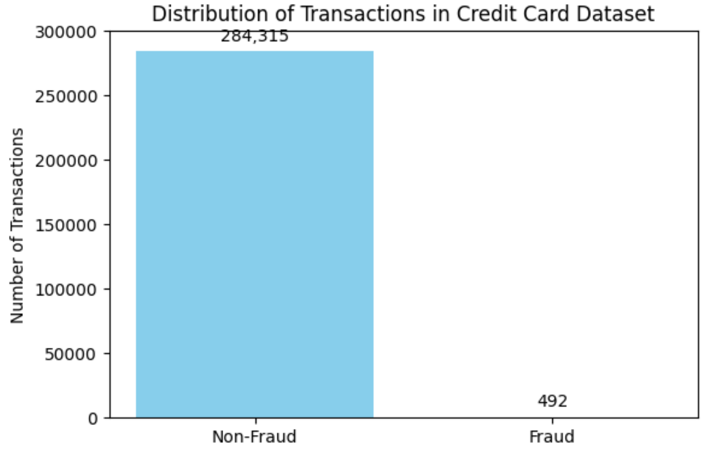


Figure 1 illustrates the extreme imbalance in the Kaggle Credit Card Fraud dataset, where only 492 out of 284,807 transactions (~0.17%) are labeled as fraud. This highlights the challenges for detection systems, as naïve models could achieve over 99% accuracy simply by predicting “non-fraud” for every transaction.

Fraud detection impacts a very wide range of stakeholders across the financial world, covering categories such as individual users all the way to global institutions. Financial institutions such as banks and credit card companies are heavily dependent on fraud detection systems that are reliably accurate to minimize losses to self and customers as well as maintain customer trust. Payment processors and online merchants are also affected, as undetected fraud can lead to chargebacks and significant reputational damage. Consumers experience the consequences of fraud detection directly through declined transactions or account freezes caused by false positives. These inconveniences can lead to "customer turnover,” where users switch to competitors after a negative security experience. Regulatory bodies also have a vested interest in ensuring that institutions implement effective safeguards against financial crime and money laundering. Fraud analysts and data scientists must balance model complexity with interpretability and operational constraints within these organizations. Poorly designed detection systems can overload analysts with alerts, reducing overall effectiveness and leading to a sort of “alert fatigue”. At the same time, overly conservative systems may fail to stop sophisticated fraud schemes that target high value assets. Comparing supervised detection with unsupervised detection methods is how this project intends to provide insights that are relevant to both technical teams and high-level decision-makers. The findings can inform the target audience on how organizations deploy data-driven algorithms for fraud detection to protect their bottom line. Ultimately, improved fraud detection benefits institutions, consumers, and the greater stability of the financial system.

Existing fraud detection solutions typically rely on supervised machine learning models trained on labeled historical data. These models can achieve strong performance when sufficient labeled examples are available for the algorithm to learn patterns. However, in fraud detection, labeled fraud cases are extremely rare and often costly or time-consuming to obtain accurately. This limitation makes supervised models vulnerable to overfitting and poor generalization when they encounter “zero-day” fraud types. Unsupervised anomaly detection methods offer an alternative by identifying unusual transaction patterns without requiring pre-existing labels. While these approaches can surface novel or emerging fraud behaviors, they may also flag many benign outliers as suspicious. Both approaches face significant challenges related to evaluation under class imbalance, where standard accuracy is a misleading metric. While we know that metrics like precision and AUPRC are vital for catching rare events, the reality of managing these systems is often much messier. Most organizations rely on a mix of hard-coded rules and machine learning, which adds a lot of moving parts to the process. Even with all the research available, we still don't have a clear answer on whether supervised or unsupervised models perform better at different scales. This study aims to fill that gap by putting both methods to the test under controlled conditions. By citing established benchmarks, we aim to clarify the boundaries between these two distinct detection paradigms.

- Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2013). *Credit Card Fraud Detection Dataset*. Kaggle. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
- Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). *A Comprehensive Survey of Data Mining-based Fraud Detection Research*. https://arxiv.org/abs/1009.6119

This project follows a structured analytical pipeline to evaluate fraud detection approaches. Before diving into modeling, we'll start with an exploratory look at the data to see how transactions are distributed and how severe the class imbalance really is. Our primary focus is the Credit Card Fraud Detection dataset, which covers about 284,000 labeled transactions. To make sure our conclusions aren't just a fluke of this specific data, we're also pulling in benchmarks from the Fraud Dataset Benchmark to validate our results across different environments. Supervised models such as logistic regression, random forests, and gradient boosting will be trained using labeled data to establish a performance baseline. In parallel, unsupervised anomaly detection techniques such as Isolation Forests or Local Outlier Factors will be applied to identify unusual transaction patterns. Model outputs will then be compared across different fraud prevalence levels and various alert thresholds to simulate real-world conditions. Evaluation will focus strictly on metrics appropriate for imbalanced data, such as recall and precision, and the area under the PR curve. The analysis is designed specifically to highlight the critical trade-offs between detection effectiveness and false positive rates. This comprehensive blueprint ensures a fair and systematic comparison of modeling approaches across different algorithmic families. The results will provide practical guidance for fraud detection system design in highly imbalanced environments.

Detecting fraud in an imbalanced environment is never straightforward. Relying on just one metric will hardly ever give you the full picture. Every team has different priorities, whether that is cutting down on financial losses or making sure legitimate customers aren't constantly getting blocked. Because of these competing goals, we've developed a series of research questions that focus on the actual trade-offs between different modeling styles. Instead of just looking at high-level totals, we're digging into transaction-level data to see how these systems hold up under real pressure. We want to understand exactly how the choice of an evaluation metric can change, or even mess up, how we perceive a model's success. By tackling these issues head-on, we hope to clarify when a supervised or unsupervised approach actually makes the most sense. Our analysis also includes a deep dive into specific features to see which ones are truly the best indicators of a fraudulent hit. These questions aren't just theoretical; they serve as a practical guide to keep the entire project grounded in real-world data science. We've designed this framework to ensure the findings are actually useful for people working in the field. Ultimately, this approach builds a solid foundation for understanding the messy reality of financial crime.

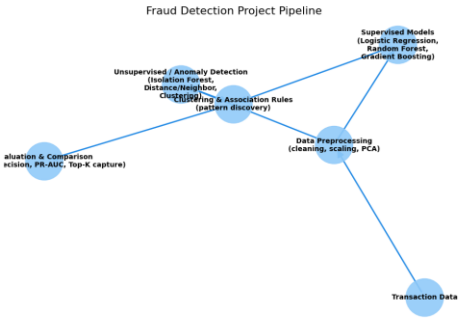
Primary Research Question

How do supervised fraud detection models compare unsupervised methods in identifying rare fraudulent transactions, and how does performance change under different fraud rates and different alert thresholds?

Supporting Research Questions

1. How do supervised models and unsupervised anomaly detectors compare at catching rare fraud as we vary the fraud rates and the "cost" of false alarms?
2. How exactly does the severity of class imbalance, meaning how rare the fraud actually is impact the performance of supervised models?
3. Does changing the fraud prevalence have a different effect on the success of unsupervised detection methods?
4. If we assume a fixed "alert budget" where a team can only check the top-K transactions, which of these two methods actually performs better?

5. What do the real world trade-offs between precision and recall look like for each approach as we shift our alert thresholds?
6. Which specific transaction features show the biggest "red flags" that allow us to separate fraud from legitimate activity?
7. Do unsupervised scores do a better job of putting real fraud at the very top of the rankings compared to supervised probabilities?
8. Are these fraud patterns and trade-offs stable enough to hold up across different datasets with varying structures?
9. Can we use clustering techniques to find specific "pockets" of transactions where fraud is much more likely to hide?
10. Under what specific conditions balancing the rarity of the fraud and the cost of an alert does one of these approaches finally outperform the other?


Fraud Detection Project Pipeline
Supervised Models (Logistic Regression, Random Forest, Gradient Boosting)
Unsupervised / Anomaly Detection (Isolation Forest, Distance/Neighbor, Clustering)
Clustering & Association Rules (pattern discovery)
Data Preprocessing (cleaning, scaling, PCA)
aluation & Comparison
ecision, PR-AUC, Top-K capture)
Transaction Data