Behavioral AI in Fraud Detection

How Sentiment, Bias & Ethics Shape the Fight Against Deception

Unified AI-Powered Fraud Detection

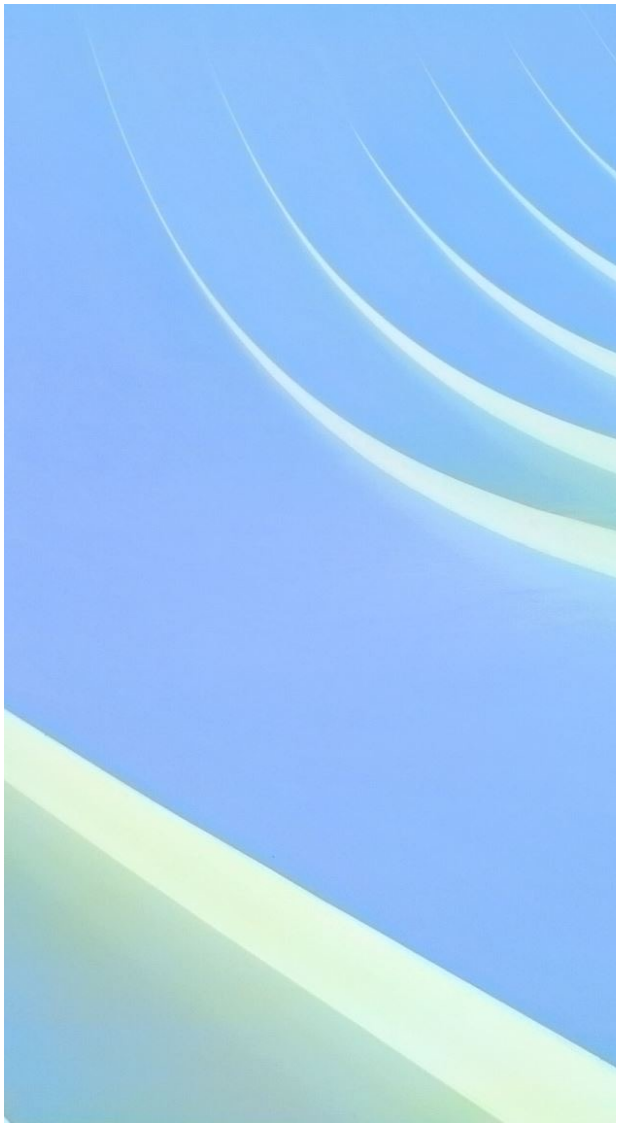
Objective: Set the stage for an interdisciplinary approach to detecting fraud.

Deep Learning (DL), a subset of machine learning powered by neural networks that mimic the human brain, rose to prominence in the 2010s and now drives advanced applications like image recognition, language translation, cancer detection, and autonomous vehicles.

Supervised learning uses labeled data to train models that can predict either discrete categories (classification) or continuous values (regression) for new, unseen inputs.

Natural Language Processing (NLP), a branch of AI focused on understanding human language, powers financial applications like sentiment analysis and SEC filing reviews and serves as the foundation for Large Language Models (LLMs) like GPT and BERT, which use deep learning to perform tasks such as text generation, summarization, and translation at scale.

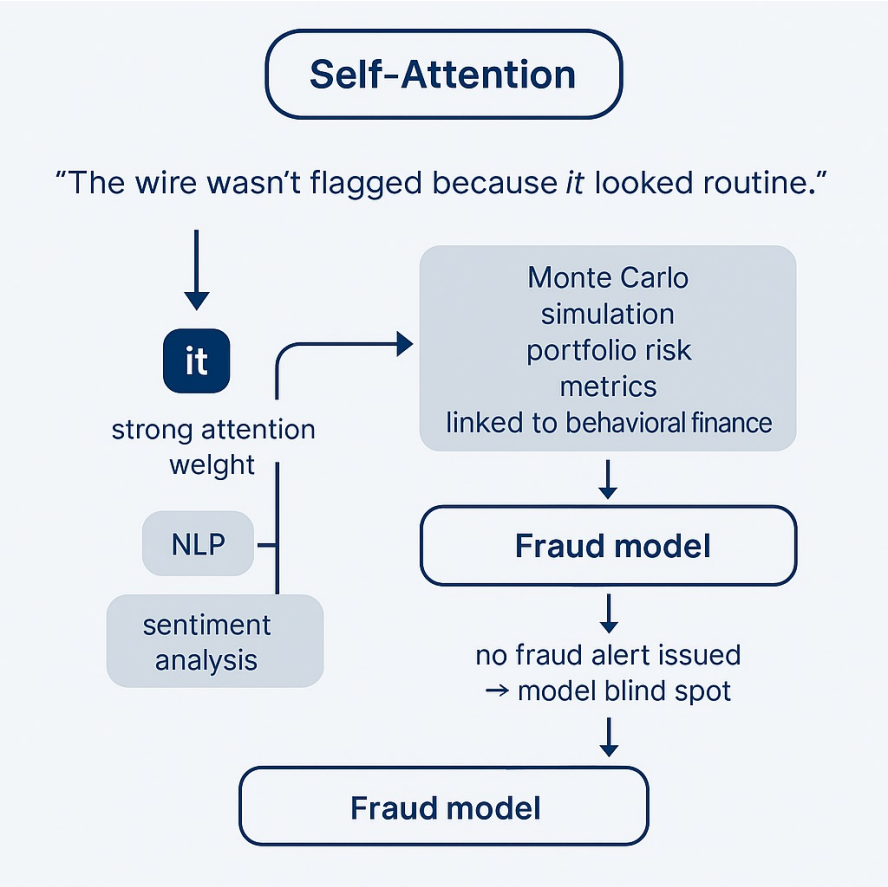
"We aim to combine behavioral insights with ML/NLP to preempt and detect fraud, not just react to it."



The Role of Self-Attention & Behavioral Finance

GPTs can be fine-tuned on financial texts to detect fraud by identifying suspicious language patterns, anomalies, and deceptive communication, making them powerful tools for early fraud detection and automated risk alerts.

Self-attention allows GPTs to trace contextual dependencies—like what "it" refers to—making them effective in decoding ambiguous or deceptive language often found in fraud-related communications.



Types Of Fraud

Objective: Categorize major fraud risks.

Types: Identity theft, insurance fraud, cyberfraud, loan fraud, insider fraud

Behavioral Angle: Overconfidence in systems leads individuals to blindly trust automated processes, ignoring subtle signs of fraud simply because "the system didn't flag it."

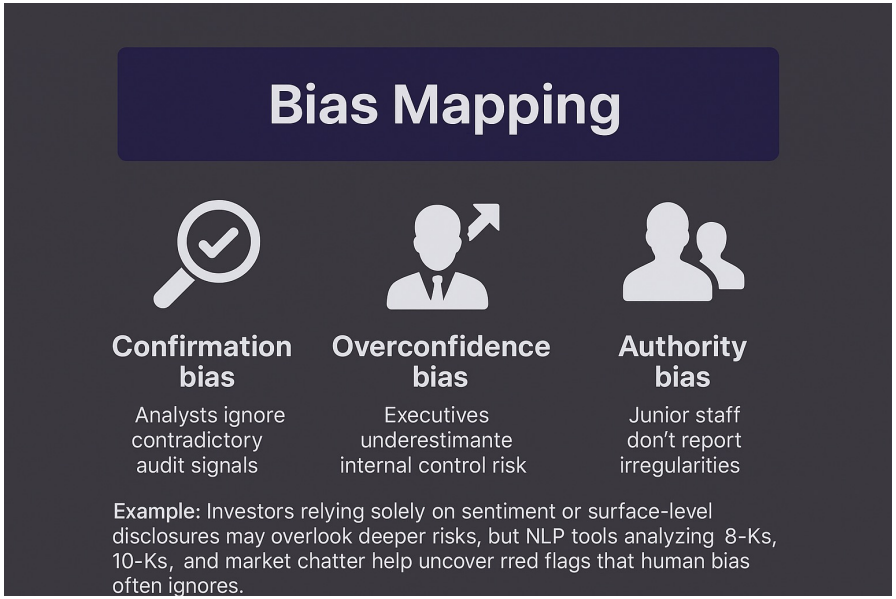
Authority bias causes employees to accept decisions from senior figures without question, enabling unethical actions to go unchecked.

Example: Text-based red flags in SEC filings—like exaggerated positivity, evasive wording, or sudden shifts in tone—can signal potential fraud and are effectively detected through advanced NLP techniques.

Identity Theft | Insurance Fraud | Cyberfraud | Loan Fraud | Insider Fraud

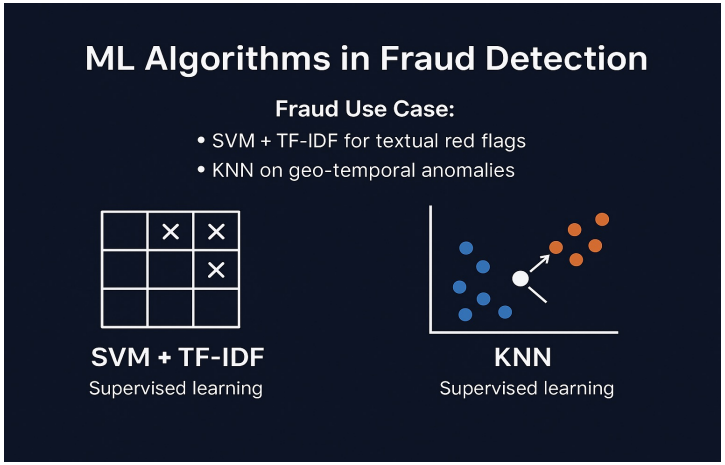
Behavioral Finance + Fraud

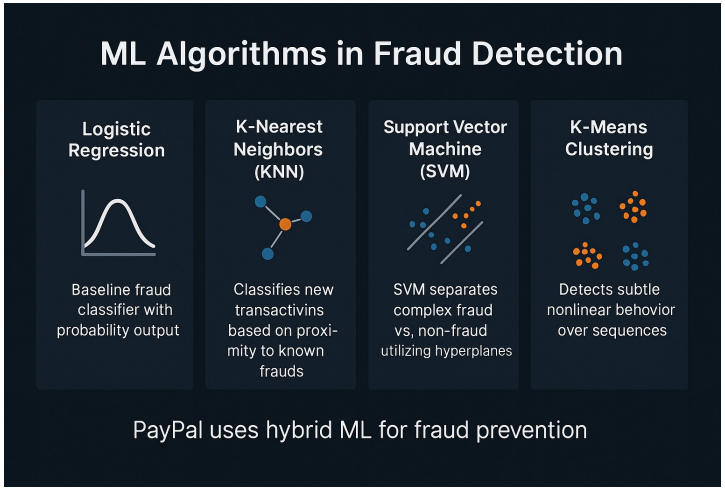
Objective: Show how human biases fuel fraud or miss it



ML Algorithms in Fraud Detection

Objective: Link ML with Fraud Flags			
Model	**Fraud Use Case**	**Strength**	**Limitation**
Logistic regression	Predicts probabilitrof fraud in credit card transactions based on structured data	Interpretable and fast	May miss complex or nonlinear fraud relations
KNN	Flags geo-temporal anomalies in payment networks by comp-aring to nearby events	Simple logic, good for local patterns	Elower with large data sets; no model memory
SVM	Classifies fraudulent invoices or claims via text and numeric fields pnihing behavior	Effective in high-dimensionl space	Requires tuning, hard to interpret

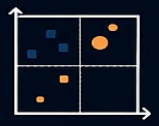




ML Algorithms in Fraud Detection

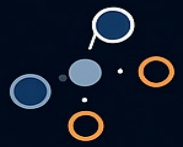
Fraud Detection Tactics Using ML Models

Use Case Strategy



SVM + TF-IDF

Textural red flags



KNN -GTIDF

Geo-temporal anomalies

Flagging suspicious patterns using proximity, freguency, and contextual indicators

ML Model Mechanics in Action

Real-World Deployment

Logistic Regression

Support Yector Machine

K-Means Clustering

PayPal uses hybrid ML combining multiple algorith ms for real-time fraud prevention

Comparative Overview: ML Models for Fraud Flag

Model Comparison Matrix

Interpretability	Speed	Fraud complexity
✓	Fast	Rasic fraud patie
⅄	Fast	Local anomalies
✕	Text 3. numerical	Teve, ζ nLon. fraud

Mapping each model to its strengths and aplication boundaries

NLP Techniques in Fraud Detection

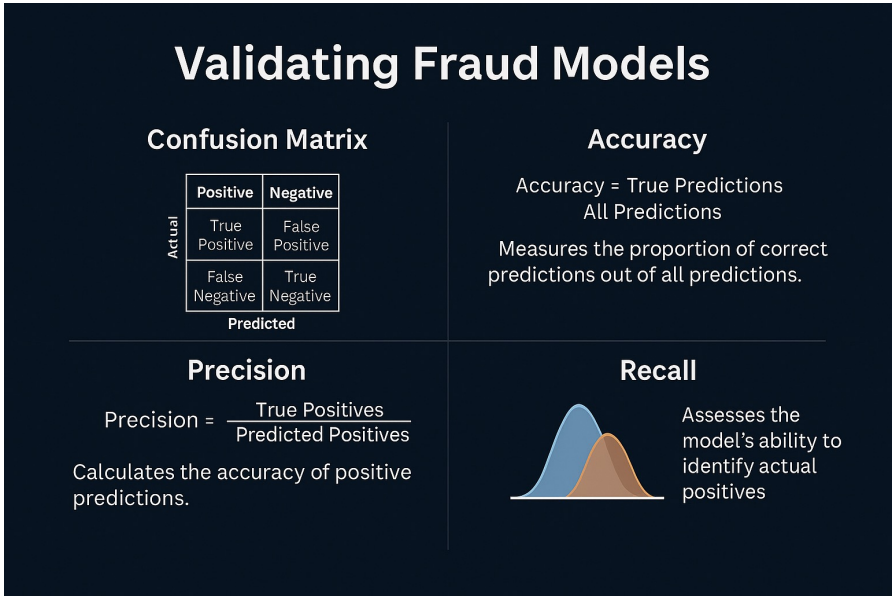
Objective: Show how text reveals deception.

TF-IDF is a word scoring method that highlights informative terms by balancing their frequency within a document against their rarity across documents, and when combined with Naïve Bayes—a simple yet powerful probabilistic classifier that assumes feature independence—it enables efficient and scalable text classification despite the naïve assumption that textual features (tokens) are uncorrelated.

In fraud detection, NLP techniques can uncover behavioral red flags by identifying frequent use of emotionally charged or deceptive language patterns—such as exaggeration or evasion—within communications or disclosures, helping flag potential fraud when combined with traditional data signals.

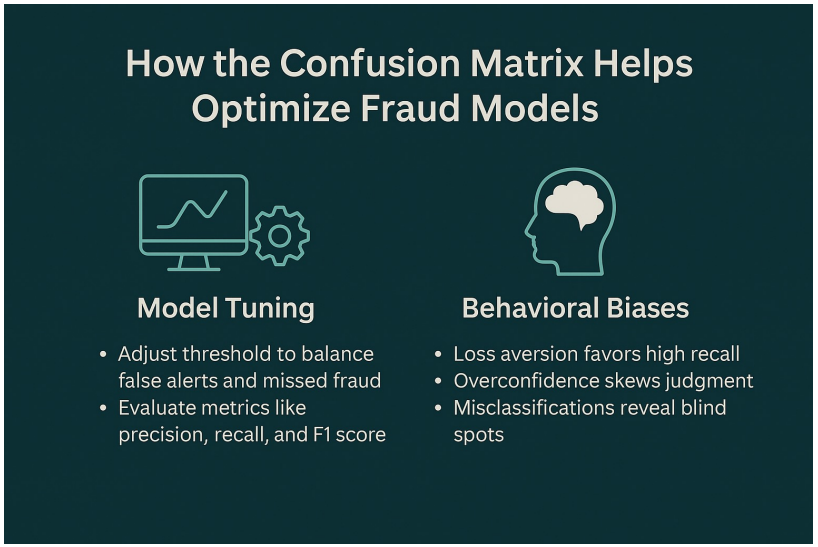
Evaluation Metrics

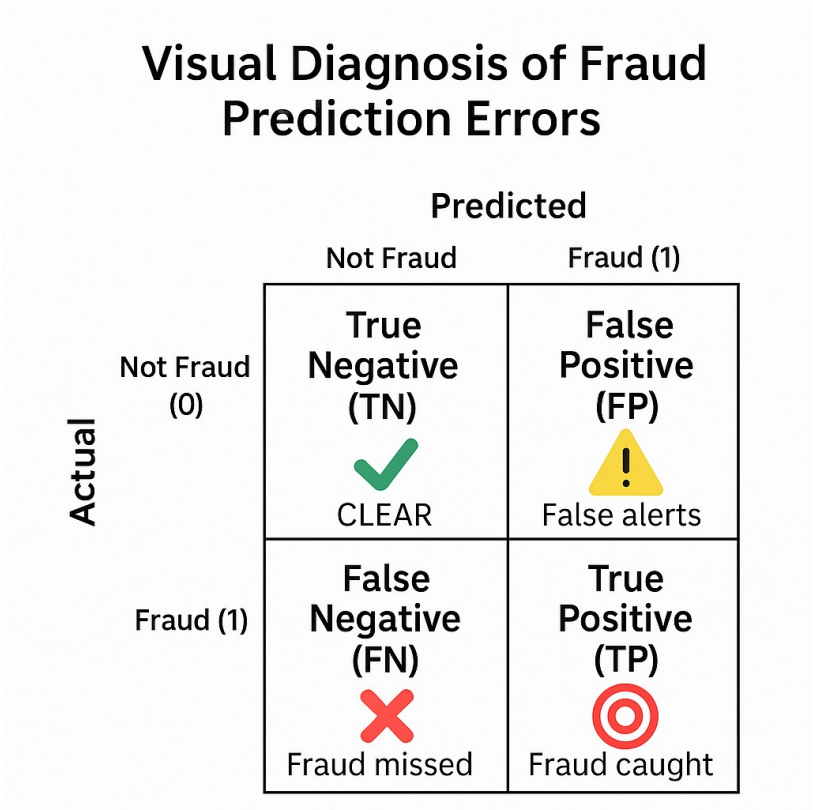
- High recall is critical to minimize undetected fraud



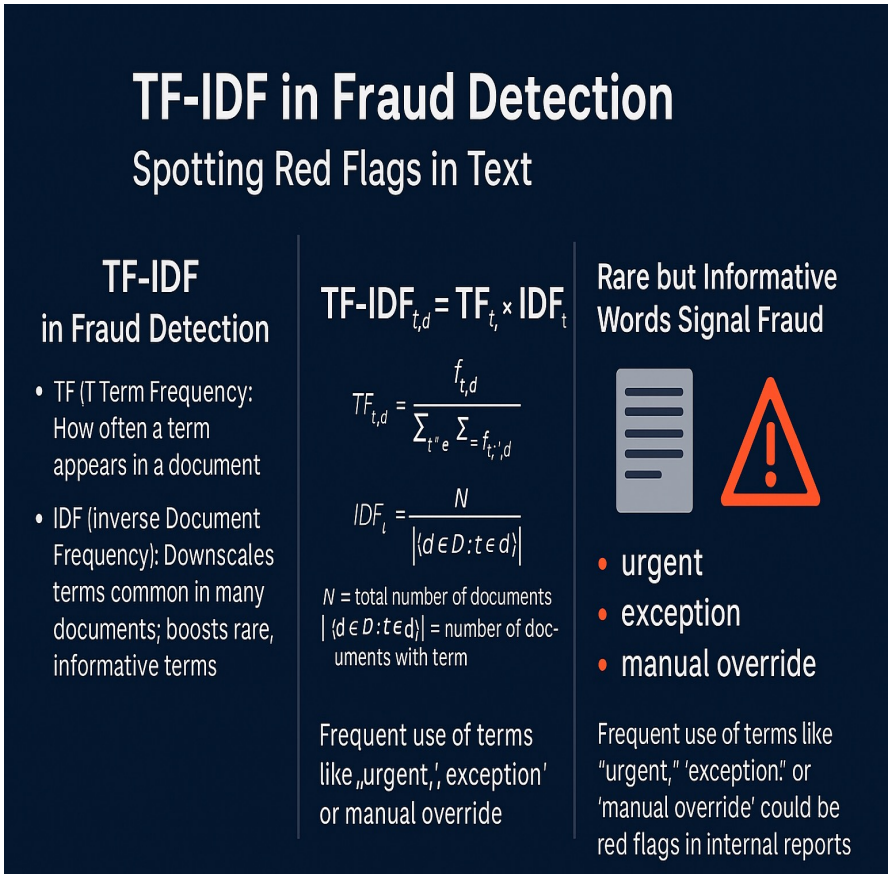
Ethics and Explainability

Confusion Matrix

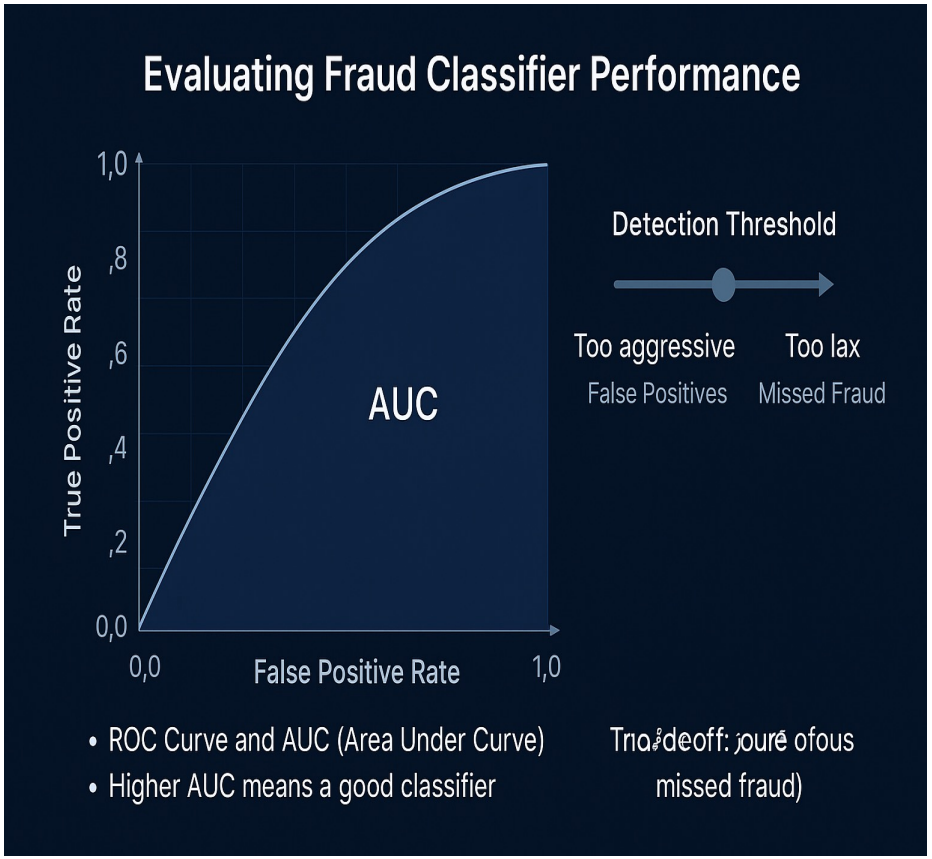




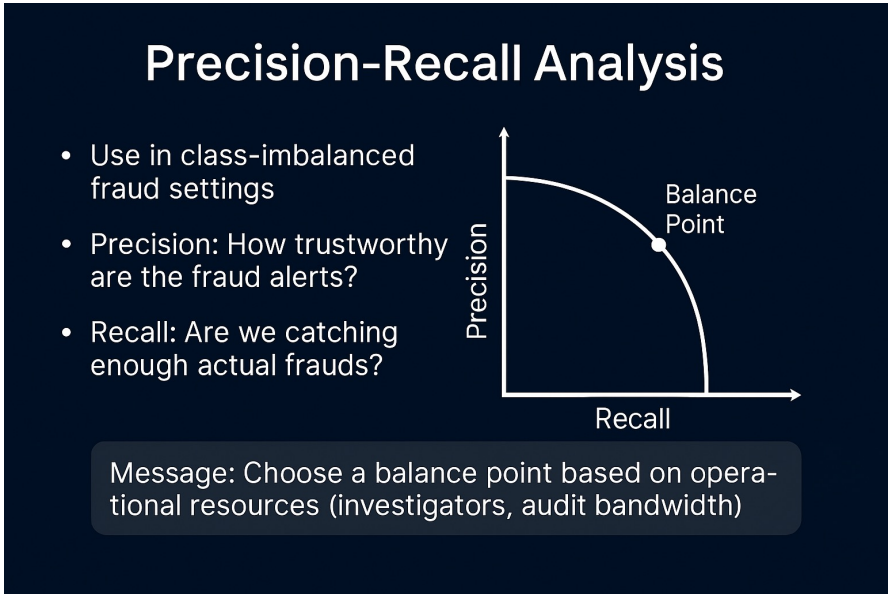
TF-IDF Heatmap



ROC Curve



Precision - Recall Curve



Key Takeaways

ML Fundamentals
Covers the basics of supervised learning in the context of fraud detection

Model Development
Explains metric evaluation, optimization, and fine-tuning of fraud classifiers

NLP Pipeline
Describes text processing and analysis steps to flag potential fraud in free-form documents

AI Risks & Safeguards
Addresses emerging risks of AI-generated fraud, emphasizing monitoring and control

Behavioral Insights
Examines human biases in decision-making and ethics in deploying fraud detection models