FraudFusion

DS 5500 – Capstone Project

Anish Rao, Raghu Ram Sattanapalle

April 15th, 2025

Presentation Overview

- Problem Specification
- Personal Learning Objectives
- Team Learning Objectives
- Related Work
- Amendments
- Solution Design
- Tool list
- Methodology
- Timeline & Key Milestones
- Sample Code
- Results
- Postmortem
- Contributions
- References

Problem Specification

Main Problem:

- Financial fraud detection is a challenging problem with extreme **class imbalance**, with fraudulent transactions making up less than 0.1% of all transactions.
- **Asymmetric misclassification costs**: Missing fraud is costlier than false positives.

Goal:

- Our main goal is to **improve the synthetic fraud** data generation process to more accurately represent the real data

Significance:

- Potentially reduce financial losses by capturing more fraudulent transactions.

Personal Learning Objectives

Anish:

- I was not really familiar with generative models before working on this project, so increasing the depth of my knowledge on this topic is my primary objective
- More experience building advanced machine learning models from scratch
- Develop expertise in designing custom loss functions for specialized domains

Raghu:

- Enhance proficiency in statistical analysis and data visualization methods to assess the quality of generated data
- Develop and refine expertise in the practical application of generative AI techniques
- Mastery in implementing and evaluating diffusion models for complex data

Team Learning Objectives

Collective Technical Goals:

- Master implementation of specialized diffusion models for tabular data with mixed feature types
- Develop a systematic approach for evaluating synthetic data quality beyond basic distribution matching
- Create reproducible methodologies for fraud detection that balance precision and recall trade-offs

Collaboration Enhancements:

- Strengthen capability to iterate rapidly through model versions with structured evaluation criteria
- Develop a shared vocabulary and framework for discussing generative model performance

Expected Takeaways:

- Comprehensive workflow for synthetic data generation in highly imbalanced domains
- Transferable approaches for custom loss function design in financial applications

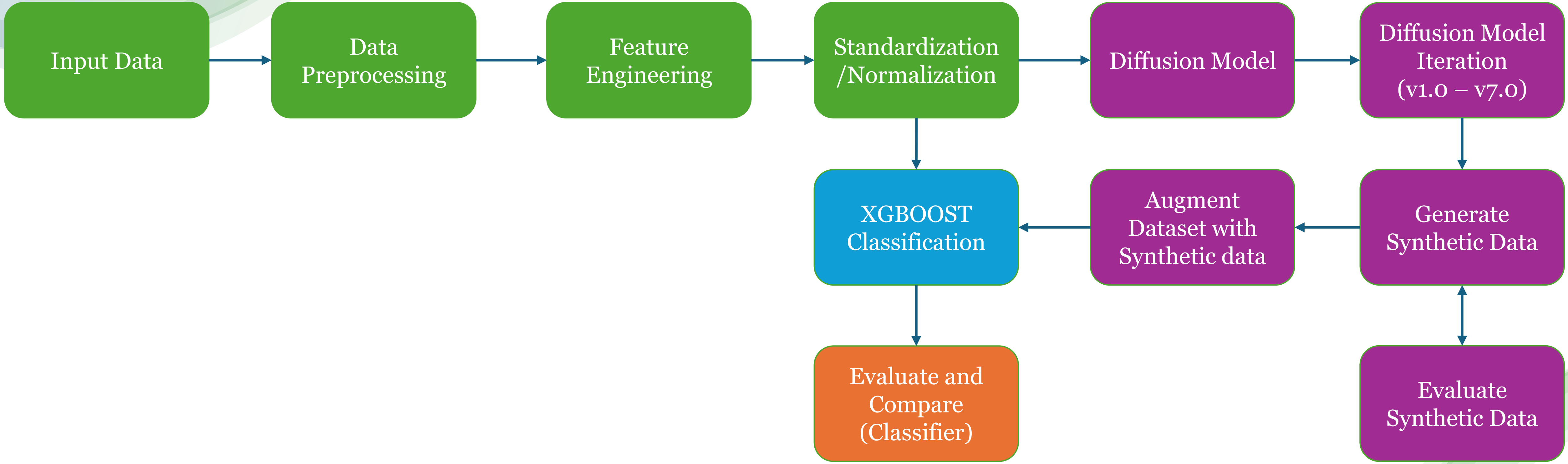
Related Work

- **FinDiff [1]:** Diffusion models for financial tabular data generation
- **Imb-FinDiff [2]:** Conditional diffusion models for class imbalance
- **TabDDPM [3]:** Diffusion models for generic tabular data
- **Dual-Track Diffusion Approach [4]:** Separate diffusion models for fraud detection
- **FraudDiffuse [5]:** Diffusion-based fraud augmentation

Amendments

- Initially we planned to work with two different financial data sets for fraud but given the decrease in manpower we decided to scale down to one data set.

Solution Design



Tool list

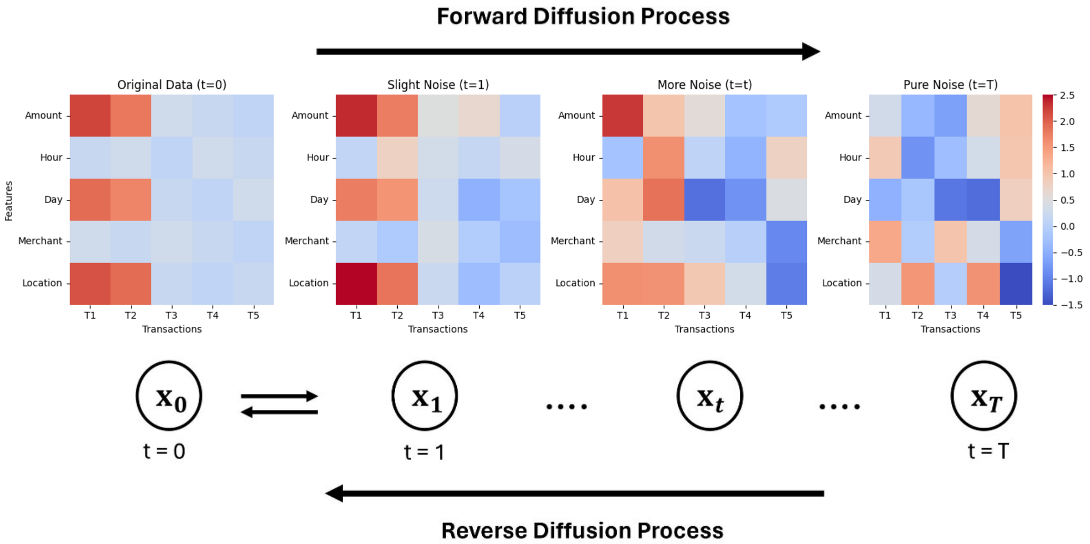
Software:

- **Programming Language:** Python 3.10
- **Deep Learning Framework:** PyTorch 2.6.0+cu118 with CUDA support
- **Machine Learning Libraries:**
  - XGBoost 2.1.3 for classification models
  - Scikit-learn 1.0.2 for preprocessing and evaluation
  - Pandas 2.2.3 and NumPy 1.26.4 for data manipulation
  - Joblib 1.4.2 for parallel processing
- **Visualization and Statistical Testing:**
  - Matplotlib 3.10.0 and Seaborn 0.13.2 for visualization
  - SciPy 1.13.1 for statistical tests (Kolmogorov-Smirnov, Anderson-Darling)
  - TQDM 4.67.1 for progress tracking during lengthy model training and generation

Hardware:

- **GPU:** NVIDIA GeForce RTX 4060 with 8GB VRAM for diffusion model training
- **CPU:** Intel Core i7 14700F (20 cores: 8P + 12E) for data preprocessing and classifier training

Methodology: Models Used



$$\boldsymbol{Forward}: x_t = \sqrt{\alpha_t}\, x_{t-1} + \sqrt{(1-\alpha_t}\,\epsilon_t$$

$$\boldsymbol{Reverse}: x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha_t}}}\epsilon_\theta(x_t, t)\right) + \sigma_t\, z$$

Enhanced Diffusion Model:

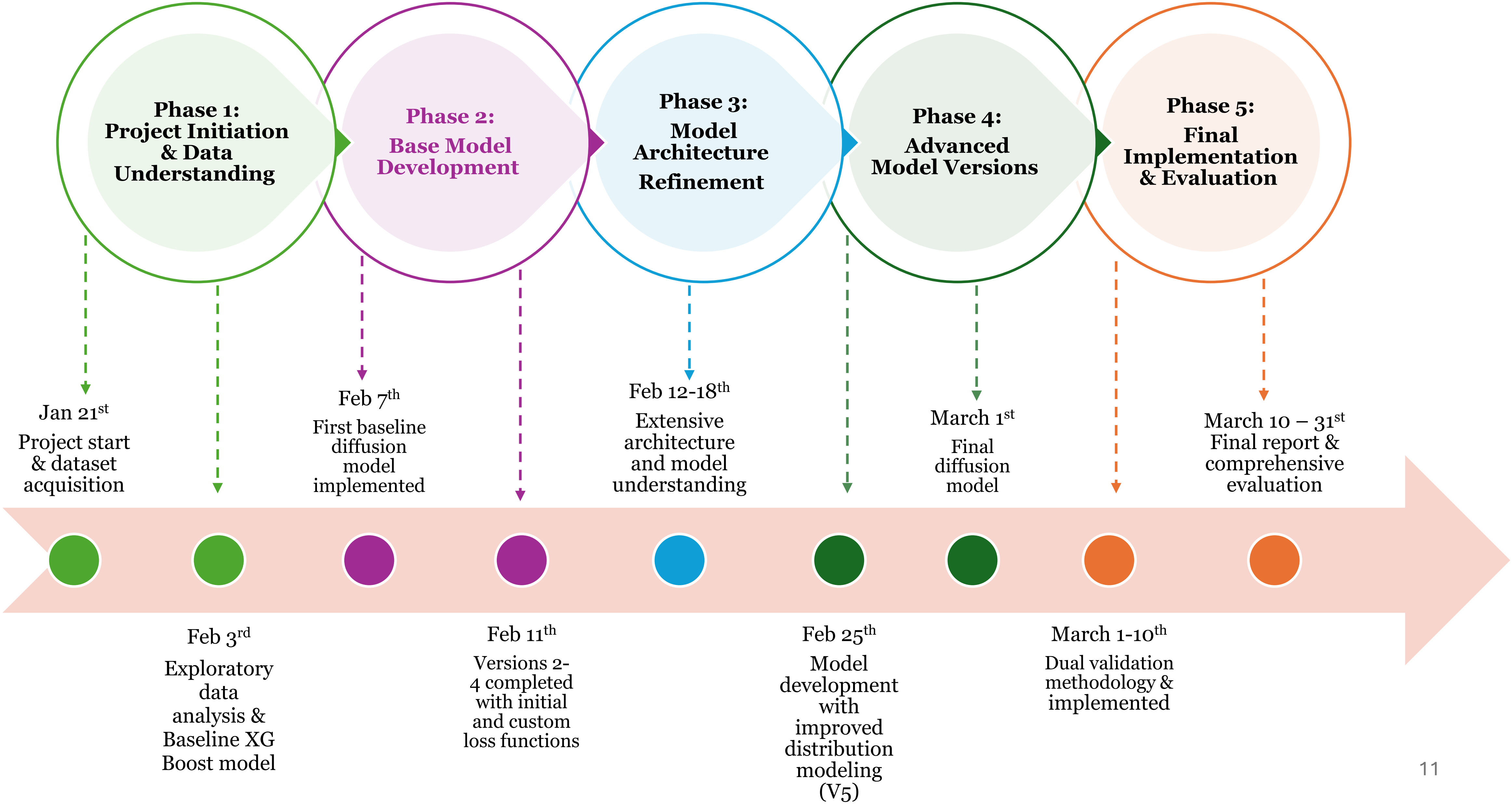
- Multi-modal neural network with feature-specific handling
- Custom loss function with 5 components for fraud-specific patterns
- Bimodal modeling for transaction amounts
- 600-step diffusion process with distribution-preserving post-processing

XGBoost Classification Models

- Baseline XGBoost model trained on original imbalanced data
- Augmented XGBoost model trained on data enhanced with synthetic fraud samples

Timeline & Key Milestones

Presenter: Raghu



Sample Video/Code

FraudDiffuse: Enhanced Diffusion Model for Synthetic Fraud Generation (v7)

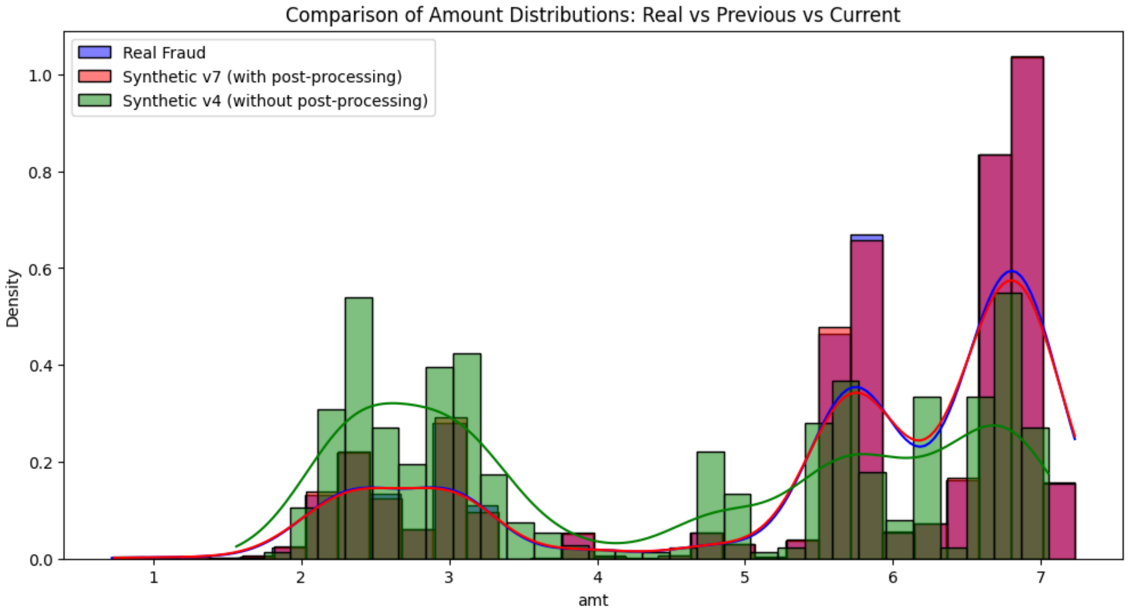
Model Overview

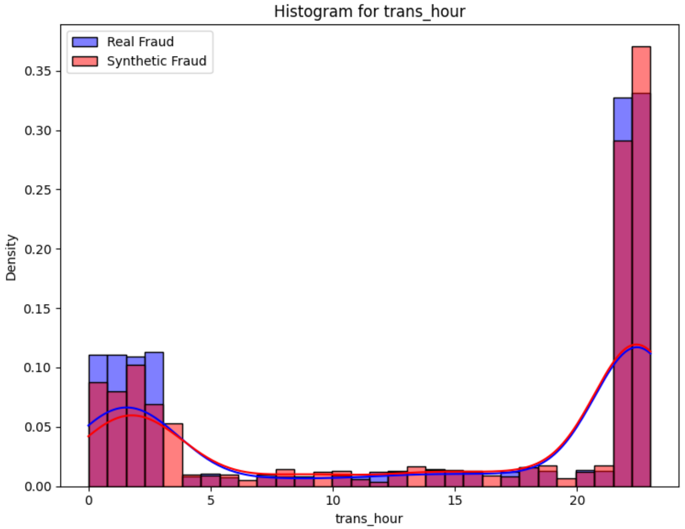
This notebook implements FraudDiffuse v7, an advanced diffusion-based generative model for creating high-quality synthetic fraud data. Key enhancements in this version include:

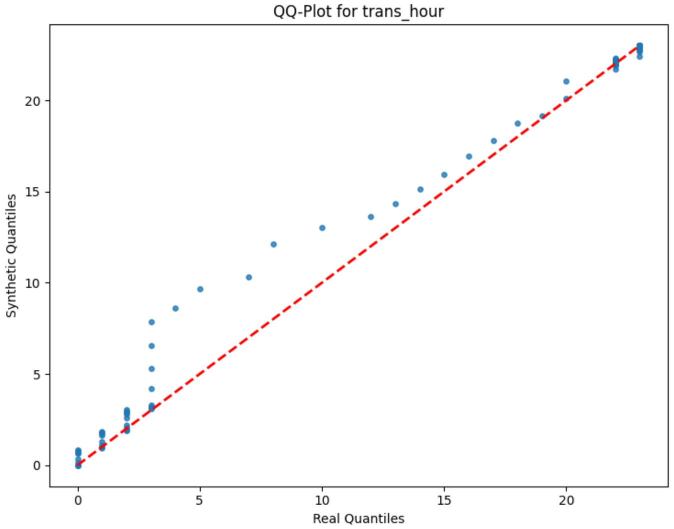
- Post-processing for amount distribution matching
- Bimodal distribution-aware initialization
- Targeted weighting for higher fraud amounts
- Distribution transformation via quantile matching

Results (Diffusion Model)

Presenter: Raghu



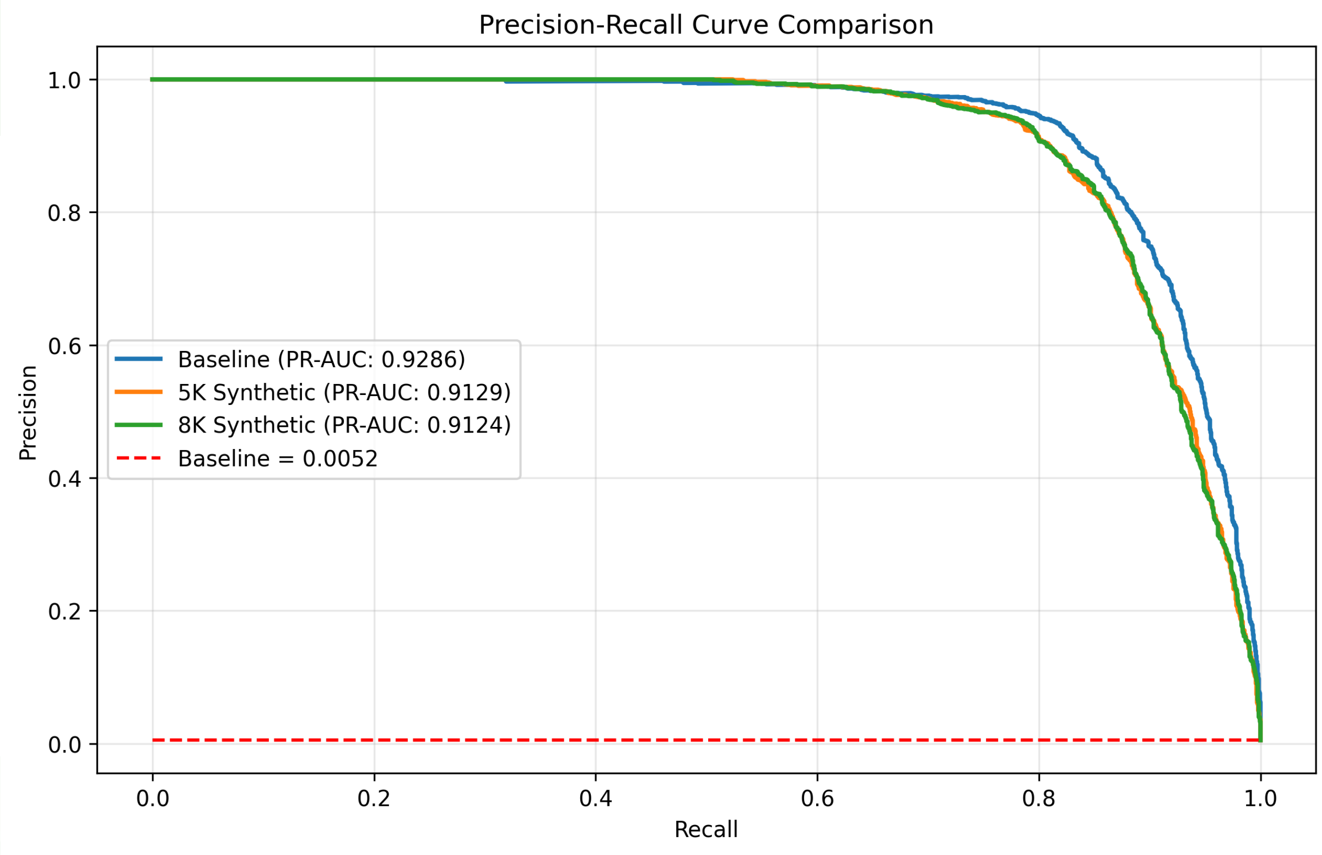




- Bimodal Amount Distribution Captured
- Version 7 (with post-processing) accurately reproduces both peaks
- Significant improvement over Version 4 (without post-processing)

- Excellent Quantile Matching (Example: 'trans_hour' feature)
- Linear trend in QQ-plot indicates synthetic data quantiles closely align with real data quantiles
- Confirms good distribution matching across features

Results (Classifier Model)



Metric	Baseline	5000 Synthetic	8000 Synthetic
ROC-AUC	0.9990	0.9984	0.9984
PR-AUC	0.9287	0.9129	0.9124
F1 Score	0.8701	0.7918	0.8069
Sensitivity/Recall	0.8275	**0.8850**	**0.8777**
Specificity	0.9173	0.9982	0.9984
Precision	0.9173	**0.7164**	**0.7466**

Postmortem

- Deepened understanding of diffusion models and their application in generating high-quality synthetic data for imbalanced classification problems
- Helped develop stronger skills in financial fraud detection and appreciate the challenges of working with highly imbalanced datasets
- Enhanced expertise in designing complex loss functions that enforce specific statistical properties in generated data
- Model improvement isn’t just about architecture changes – sometimes targeted distribution matching and feature specific optimizations make the biggest difference
- Taught us importance of rigorous validation methodologies when working with synthetic data

Future Direction

Model Enhancements:

- **Temporal Modeling:** Incorporate sequence modeling for transaction streams to detect complex fraud patterns over time
- **Transfer Learning:** Pre-train diffusion models on larger financial datasets, then fine – tune for specific types

Expanded Applications:

- **Cross-Domain:** Apply to insurance fraud, anti-money laundering, and other financial crimes
- **Multimodal Fusion:** Combine transaction data with text (e.g., transaction descriptions)

Further Research:

- **Adaptive Generation:** Design systems that dynamically adjust synthetic data characteristics based on classifier feedback
- **Advanced Distribution Matching:** Explore adversarial distribution matching for even more realistic synthetic data

Contributions (Raghu)

Feature Engineering:

- Developed cyclic encoding for temporal fraud patterns

Diffusion Model development & testing:

- Developed & enhanced FraudDiffuse architecture with custom components for mixed data types
- Designed & implemented custom loss functions (Amount distribution & Feature weighted MSE loss)
- Created framework for synthetic data quality assessment
- Implemented specialized bimodal distribution matching for transaction amounts
- Implemented post –processing distribution transformation techniques

Contributions (Anish)

Feature Engineering & Data Preprocessing:

- Developed comprehensive **pipeline** for cleaning and **preparing transaction data** for diffusion model iterations
- Created specialized transformations for **temporal and categorical** features

Diffusion Model development & testing:

- Architecture & Hyper-parameter testing for diffusion model improvements

XGBoost Implementation:

- Designed and optimized the fraud detection classifier pipeline
- Created controlled synthetic data augmentation framework
- Conducted performance analysis across precision-recall spectrum

References

1. Sattarov, T., Schreyer, M., & Borth, D. (2023). FinDiff: Diffusion models for financial tabular data generation. Proceedings of the 4th ACM International Conference on AI in Finance.
2. Schreyer, M., Sattarov, T., Sim, A., & Wu, K. (2024). Imb-FinDiff: Conditional diffusion models for class imbalance synthesis of financial tabular data. Proceedings of the 5th ACM International Conference on AI in Finance.
3. Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023). TabDDPM: Modelling tabular data with diffusion models. Journal of Machine Learning Research.
4. Pushkarenko, N., & Zaslavskyi, V. (2024). Synthetic data generation for fraud detection using diffusion models. Financial Technology and Machine Learning.
5. Roy, R., Tiwari, D., & Pandey, A. (2023). FraudDiffuse: Diffusion-aided synthetic fraud augmentation for improved fraud detection. arXiv preprint.

Questions?

Thank you!