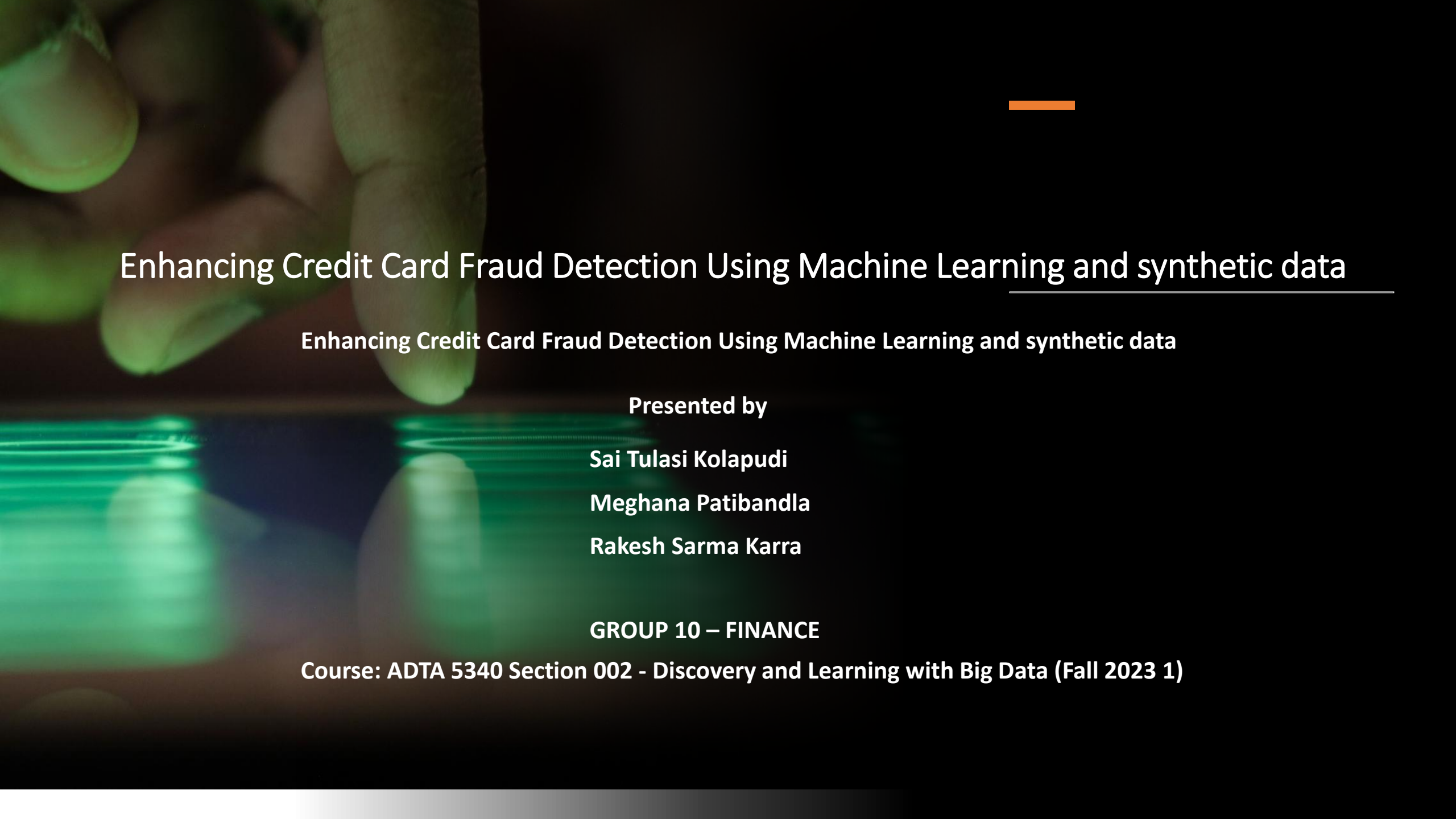


Enhancing Credit Card Fraud Detection Using Machine Learning and synthetic data

Enhancing Credit Card Fraud Detection Using Machine Learning and synthetic data

Presented by

Sai Tulasi Kolapudi

Meghana Patibandla

Rakesh Sarma Karra

GROUP 10 – FINANCE

Course: ADTA 5340 Section 002 - Discovery and Learning with Big Data (Fall 2023 1)

Agenda

Navigating Through Our Approach to Credit Card Fraud Detection

1. Introduction to Credit Card Fraud
2. The Challenge of Data Imbalance
3. Overview of Dataset and Preprocessing
4. Role of GANs in Data Augmentation
5. Exploratory Data Analysis (EDA)
6. Machine Learning Models Employed
7. Model Evaluation Metrics
8. Implications and Deployment Strategies
9. Conclusion and Future Work
10. Question & Answer Session

Introduction to Credit Card Fraud

The Growing Challenge

"Credit card fraud, a significant issue in digital finance, presents complex challenges due to the rise in online transactions."

Machine Learning as a Solution

"Traditional detection methods fall short against sophisticated fraud techniques, highlighting the need for advanced machine learning solutions."

Our Focus

"Our project utilizes a data-driven approach, employing machine learning enhanced with synthetic data, to effectively identify fraudulent activities in a highly imbalanced dataset."

Problem Statement

Data Imbalance Issue

"Tackling the challenge of detecting rare fraudulent transactions (492 out of 284,807) in a highly imbalanced dataset.“

Privacy and Data Utility

"Addressing privacy concerns by utilizing synthetic data generation to replicate real transaction patterns without compromising customer data confidentiality.“

Objective

"Aim to develop a balanced, privacy-conscious model with enhanced accuracy in fraud detection, reducing false positives."

Overview of Dataset and Preprocessing

Dataset Overview

- "Analyzed a dataset containing 284,807 credit card transactions, of which 492 are fraudulent, from September 2013 involving European cardholders."
- "Features include 28 anonymized variables from PCA, along with 'Time' and 'Amount', ensuring data privacy.“

Initial Data Insights

- "Data attributes (V1-V28) are PCA-transformed, with 'Time' representing seconds since the first transaction and 'Amount' indicating transaction value."
- "The dataset shows no missing values across all columns, indicating completeness."

Preprocessing Techniques

- "Standardization: Applied to 'Amount' to normalize transaction values."
- "Time Conversion: Transformed 'Time' from seconds to hours to capture transaction trends over the day."
- "Outlier Analysis: Evaluated outliers within fraudulent and non-fraudulent transactions, noting a higher fraction of outliers in fraud cases (e.g., V10, V12, V14).“

Balancing Privacy and Anomaly Detection

- "PCA transformation maintains user privacy. Outlier handling was cautiously approached, as outliers could represent genuine fraud cases."

Objective of Preprocessing

- "Our preprocessing aims to optimize the dataset for machine learning, ensuring data quality and representative features for detecting fraud patterns."

Role of GANs in Data Augmentation

From Imbalance to Equilibrium

"Originally, our dataset exhibited a severe imbalance with 284,315 non-fraudulent and only 492 fraudulent transactions. We've successfully transformed it using GANs."

"Post-augmentation, the dataset now encompasses over 500,000 transactions, balancing the scales at 400,000 non-fraudulent and 100,000 fraudulent transactions."

Role of GAN in Data Enrichment

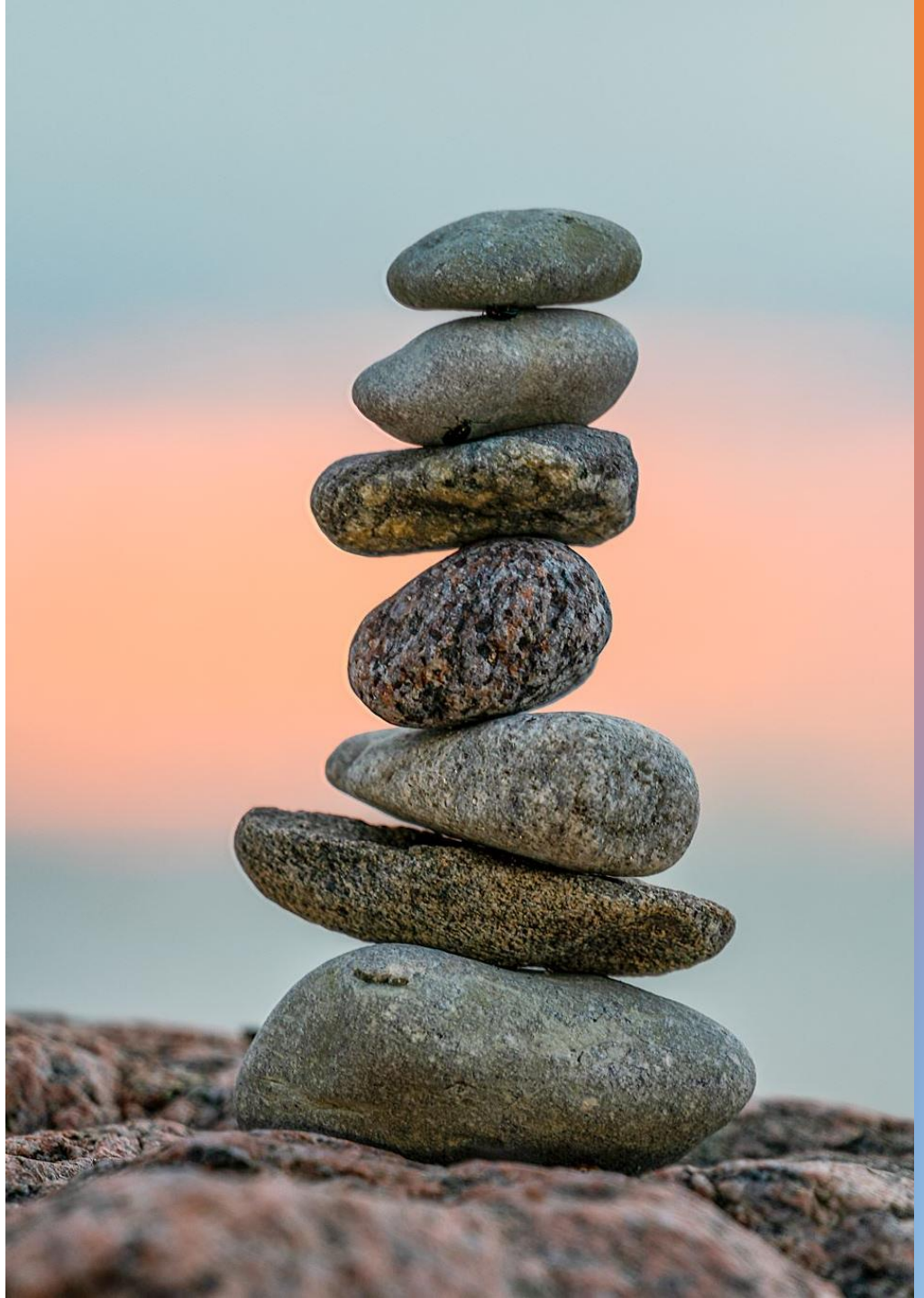
"The GAN's Generator synthesized realistic fraud samples, while the Discriminator refined their quality, ensuring the synthetic data closely mimics actual fraud patterns."

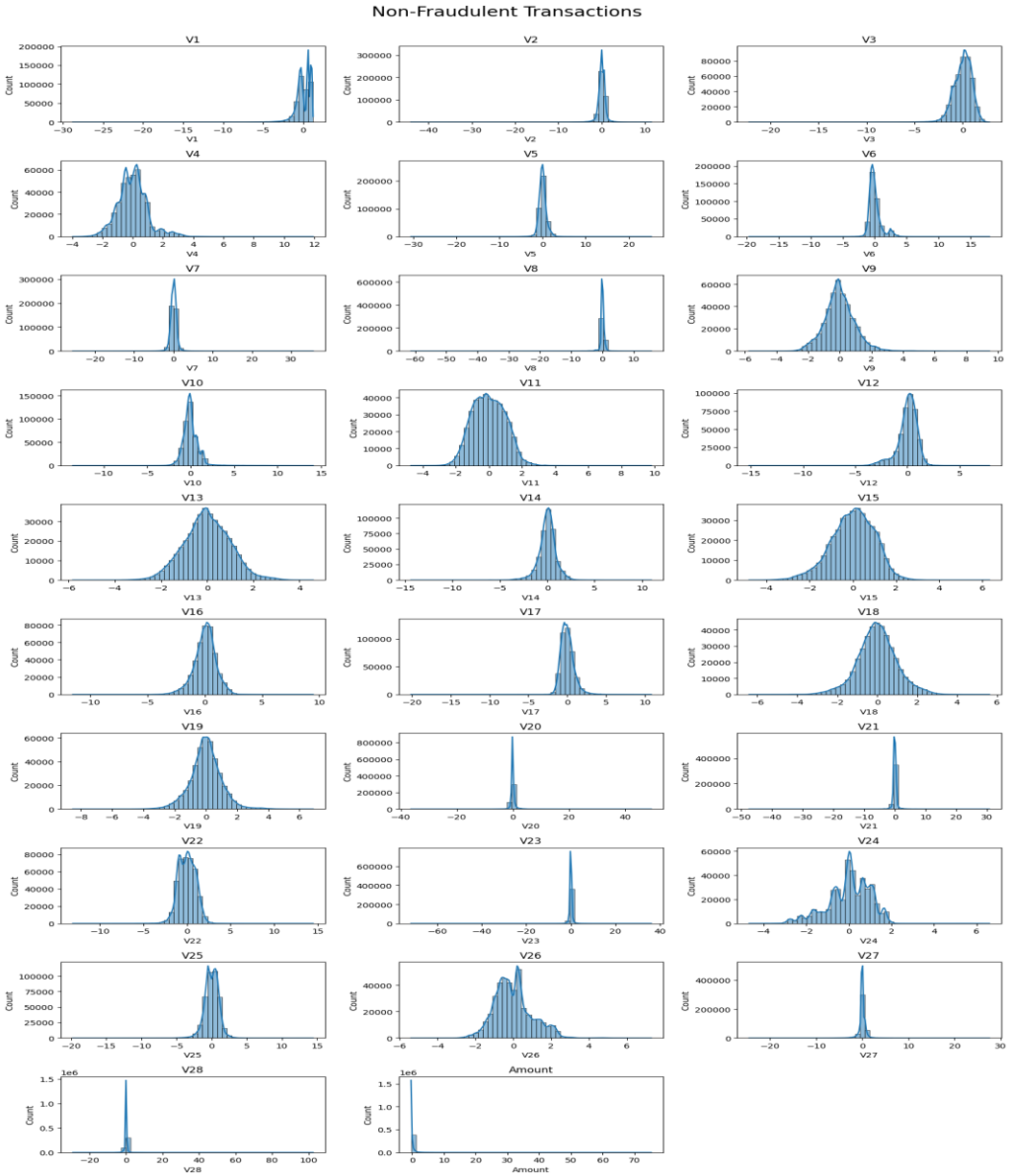
"This process was critical in creating a diverse and representative dataset for robust model training."

Impact on Machine Learning Models

"The enriched dataset, featuring a substantial increase in fraudulent cases, provides a solid foundation for training more effective fraud detection models."

"This approach not only addresses the class imbalance but also upholds data privacy, minimizing reliance on sensitive real-world data."





Exploratory Data Analysis (EDA)

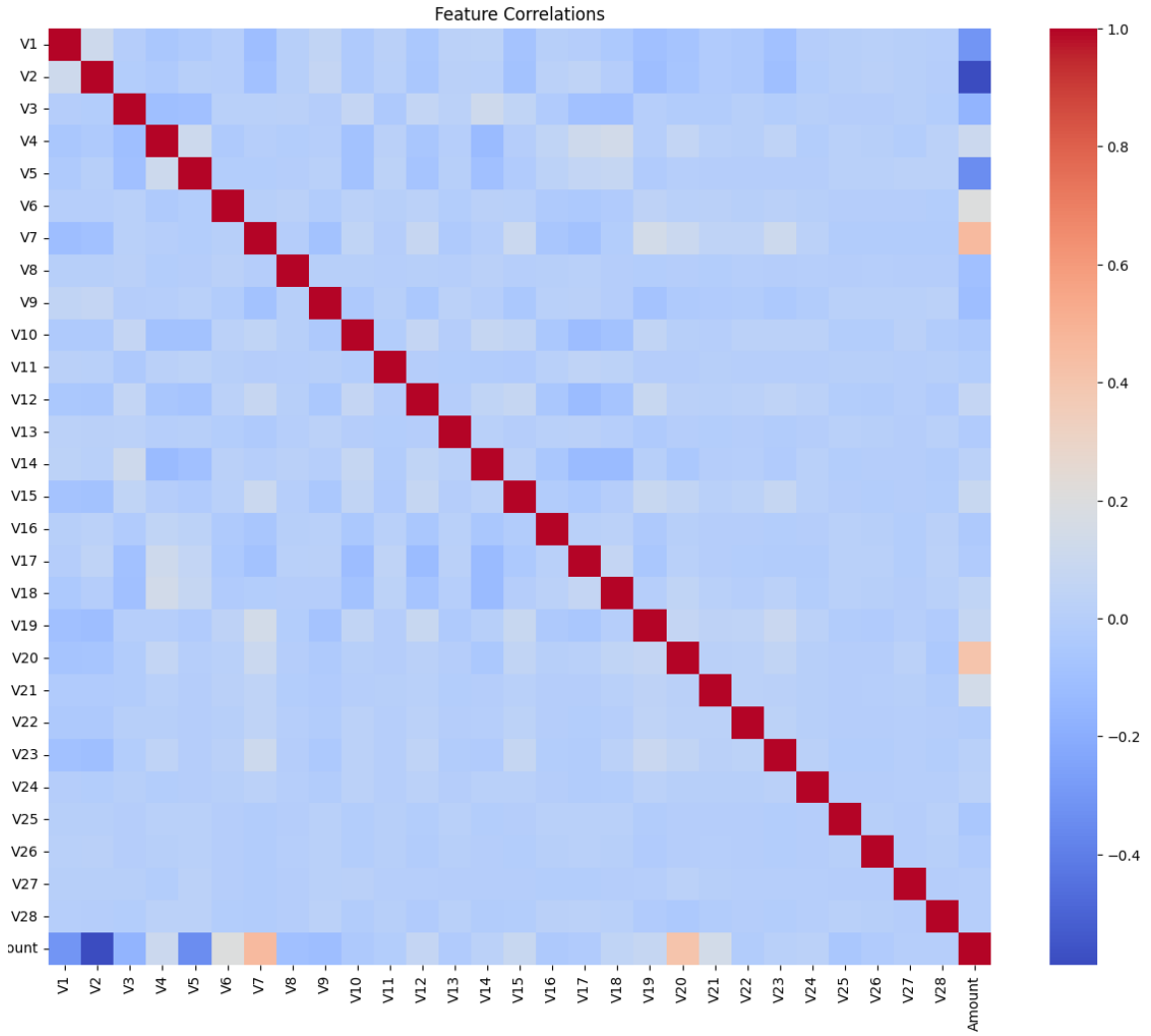
Feature Distribution Analysis

"Histograms indicate a normal distribution for V5, V6, and V26 among non-fraudulent transactions, contrasting with the skewed distributions in fraudulent ones.

"V4 and V11 show a higher prevalence of extreme values in fraudulent transactions, suggesting their potential as strong predictors of fraud."

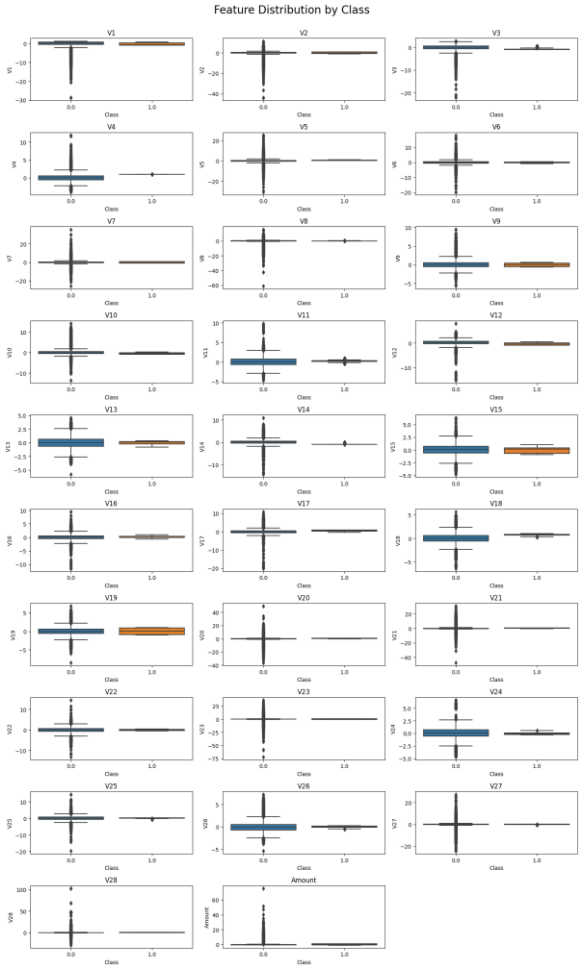
Correlation Heatmap Findings

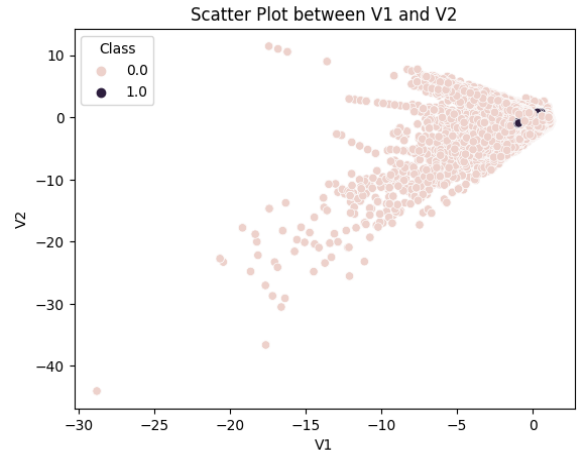
- "The heatmap reveals minimal correlation between PCA features, with a few exceptions such as V2 and V5, indicating potential multicollinearity.
- "The Amount feature shows low correlation with PCA features, confirming its independence and potential value in the model."



Scatter and Box Plot Observations

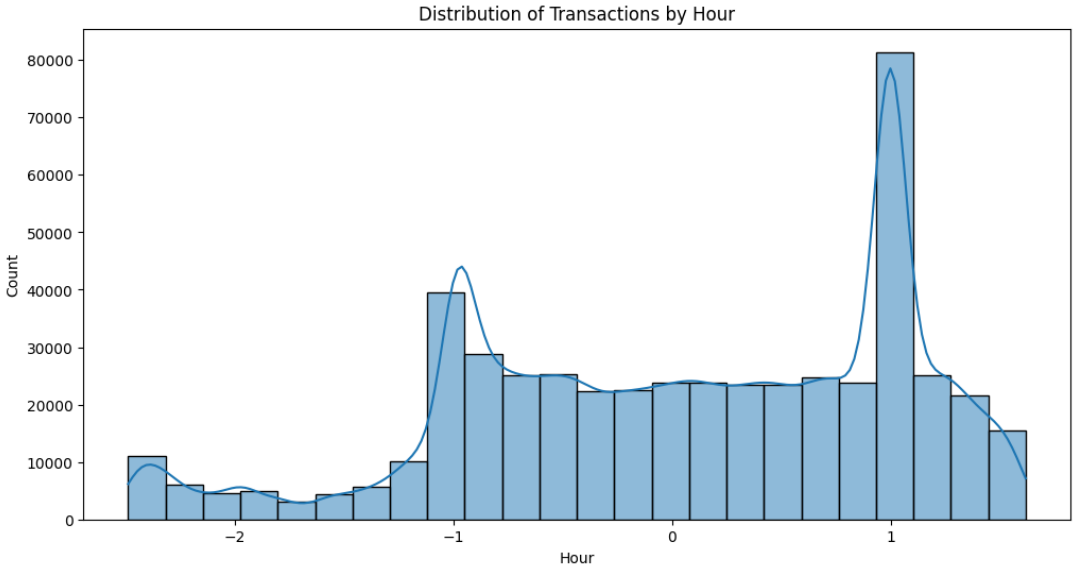
- "Scatter plots for V1 and V2 show a distinct clustering of fraudulent transactions, differentiating them from non-fraudulent ones"
- "Box plots exhibit wider interquartile ranges and more extreme outliers for fraud cases in features like V7 and V10, which could be indicative of fraudulent activity.

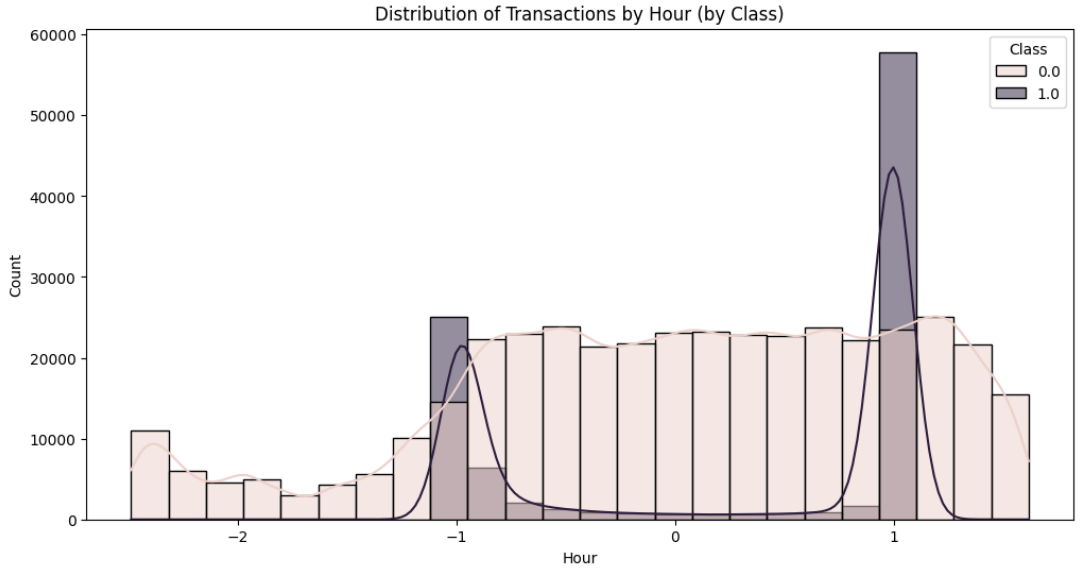




Temporal Pattern Evaluation

- "Hourly transaction distribution unveils a bimodal pattern, peaking around the typical hours of human activity"
- "Fraudulent transactions do not follow this bimodal distribution, suggesting atypical activity times for fraud occurrences."





Machine Learning Models Employed

Data Split for Model Training:

- "Employed an 80-20 train-test split on our balanced dataset, resulting in 399,606 training samples and 99,902 for testing, primed for model validation."

Diverse Model Deployment:

- "Logistic Regression for baseline probability assessments."
- "Random Forest and XGBoost for their powerful ensemble learning capabilities."
- "KNN and Gaussian Naive Bayes for pattern recognition and probabilistic classification."
- "SVM for high-dimensional feature handling."
- "Decision Tree for clear decision pathways and interpretability."

Model Evaluation Metrics

Metric Selection

"Our models were evaluated using a range of metrics crucial for imbalanced classes, including AUPRC, accuracy, recall,and precision.

Precision-Recall Tradeoff

"AUPRC is especially informative for imbalanced datasets, focusing on the precision-recall tradeoff, which is vital for fraud detection applications."

Insights

- "The table summarizes the key evaluation metrics for each model, highlighting their performance in fraud detection."
- "Models like Random Forest, XGBoost, KNN, and Decision Trees achieved perfect scores across multiple metrics, indicating robustness and effectiveness."
- "The evaluation metrics provide confidence in our models' ability to accurately detect fraud, with the potential to significantly reduce false positives and false negatives in real-world scenarios."

Model	AUPRC	Accuracy	Recall	Precision
Logistic Regression	0.8687	0.9709	0.9573	0.9024
Random Forest	1.0000	1.0000	1.0000	1.0000
XGBoost	1.0000	1.0000	1.0000	1.0000
K-Nearest Neighbors	1.0000	1.0000	1.0000	1.0000
Naive Bayes	1.0000	0.9998	0.9990	1.0000
Support Vector Machine	-	1.0000	1.0000	-
Decision Trees	1.0000	1.0000	1.0000	1.0000

Implications and Deployment Strategies

Model Excellence:

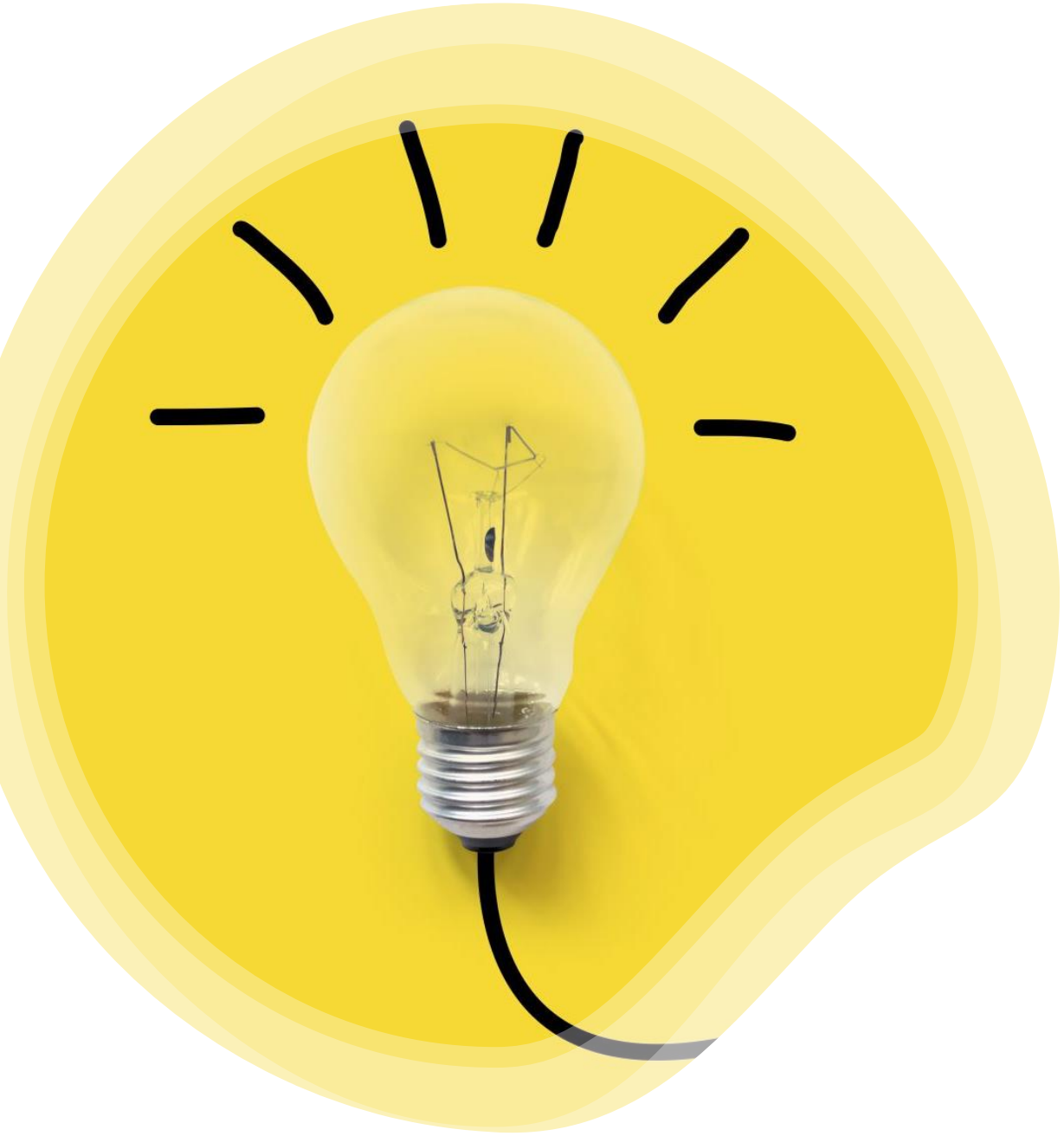
- "Our models achieve over 99% accuracy, providing real-time protection against financial fraud."

Effective Deployment:

- "Seamlessly integrate models into transaction systems for immediate fraud identification."
- "Use real-time alerts for swift action."

Resource Optimization:

- "Leverage cloud-based scalability and cost-efficiency."
- "Minimize latency for high-speed transactions."



Human Collaboration

- "Engage experts for complex cases and decision validation."
- "Foster AI-human collaboration."

Ethical Considerations

- "Assess model fairness to avoid bias."
- "Comply with data privacy regulations."

Customer-Centric

- "Design user-friendly alerts."
- "Prioritize customer experience without compromising security."

Continuous Improvement

- "Establish feedback loops for ongoing model enhancement."
- "Monitor performance and gather feedback."

Conclusion

Strategic Machine Learning Integration: Effectively leveraged diverse machine learning models, achieving significant strides in credit card fraud detection.

Innovative Data Handling:

- **Robust Data Analysis & Preprocessing:** Focused on feature standardization and comprehensive outlier analysis to spotlight fraudulent transactions.
- **Data Privacy with GANs:** Utilized Generative Adversarial Networks to generate synthetic data, enhancing privacy and addressing imbalances in the dataset.

Model Performance Breakthroughs:

- **Realistic Insights:** Logistic Regression provided balanced detection outcomes.
- **Exceptional Accuracy:** Achieved near-perfect metrics with Random Forest, XGBoost, KNN, Naive Bayes, SVM, and Decision Trees, though mindful of overfitting risks.

Core Accomplishments:

- Advanced nuanced understanding of fraud detection mechanisms.
- Emphasized the importance of data balance and privacy, pioneering the use of synthetic data in fraud detection.

Future Work

- **Model Validation and Testing:** Further validation on different datasets and using techniques like cross-validation is necessary to confirm the generalizability and robustness of the models.
- **Continuous Model Update:** Ongoing updating and retraining of models with new transaction data will be crucial to adapt to evolving fraud patterns.
- **Feature Importance and Explainability:** Future work will delve into understanding which features most significantly contribute to predictions and enhancing the explainability of the models, which is vital in sensitive applications like fraud detection.
- **Real-Time Implementation:** Efforts will be directed towards integrating these models into real-time fraud detection systems, ensuring prompt and efficient detection and response.
- **Ethical and Regulatory Compliance:** Ensuring adherence to ethical guidelines and data privacy regulations will remain a cornerstone of future developments in this area.

Thank You!

Any Questions?