Conquering Class Imbalance: Strategic Sampling for High-Recall Credit Card Fraud Detection

A Machine Learning Approach

Presenter: Shayma Remy

Springboard | Data Science Track – Capstone | June 2025

Executive Summary - The Challenge & Our Solution

- **The Problem:** Credit card fraud costs billions annually; 3.5% of transactions in our dataset, but high cost of missed fraud.
- **Our Goal:** Develop a ML pipeline prioritizing **recall** (detecting fraud) while maintaining acceptable **precision**.
- **Core Strategy:** Addressed extreme class imbalance (20k fraud vs. 570k legitimate) and large data volume (1.9 GB).
- **Key Approach:** 2:1 undersampled subset + SMOTE for balanced training.
- **Result:** Random Forest model with 73.34% recall, 85.02% precision (after threshold tuning), 0.9176 ROC AUC.

Introduction - The Scale of Credit Card Fraud

Why Fraud Detection Matters: Billions in annual losses, customer trust erosion, regulatory scrutiny.

The Data Landscape:

Merged IEEE-CIS dataset: 590,540 transactions, 144,233 identity records.	Extreme Class Imbalance: Only 3.5% (20,663) fraudulent transactions.	High Data Volume: Approx. 1.9 GB in memory, demanding efficient processing.

The Core Problem: Naive models fail; predicting "legitimate" yields high accuracy but misses virtually all fraud (near zero recall).

Data Overview & Preprocessing Steps

Dataset Merging: Joined *train_transaction.csv* (394 cols) and *train_identity.csv* (41 cols) on TransactionID, resulting in 590,540 rows and 434 columns.

Initial Assessment: Identified high missing rates in features like dist2, D7, DeviceInfo.

Memory Management: Reduced RAM usage from 1.9 GB to 1.8 GB via dtype optimization.

Preprocessing Pipeline:

- Missing Value Imputation: Numeric (median), Categorical ("Unknown").
- Feature Engineering: Extracted temporal features (hour, weekday) from TransactionDT.
- Encoding & Scaling: RobustScaler for numeric, Label/One-Hot Encoding for categorical features.

Strategic Training: Undersampling & SMOTE

The Challenge with Full Dataset Training:

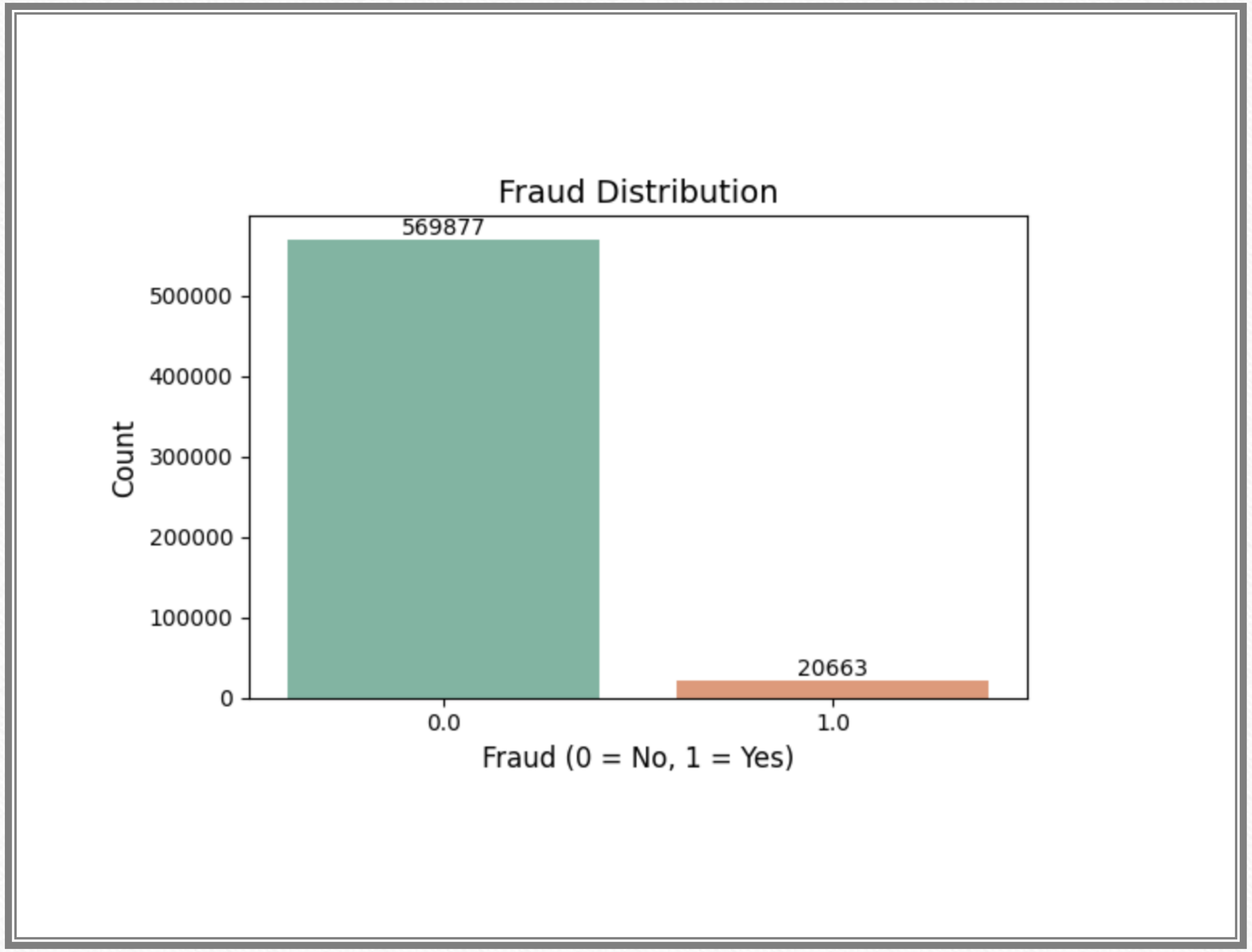
- Hours of training time required.
- Sophisticated imbalance handling needed.
- Typically, low recall due to overwhelming legitimate transactions.

Our Chosen Approach: 2:1 Undersampled + SMOTE:

- Retained all 20,663 fraud cases.
- Randomly sampled 41,326 legitimate cases (2:1 ratio).
- Resulted in 61,989 rows (33.3% fraud).
- Applied SMOTE to training set to achieve 50/50 class balance (66,122 rows).

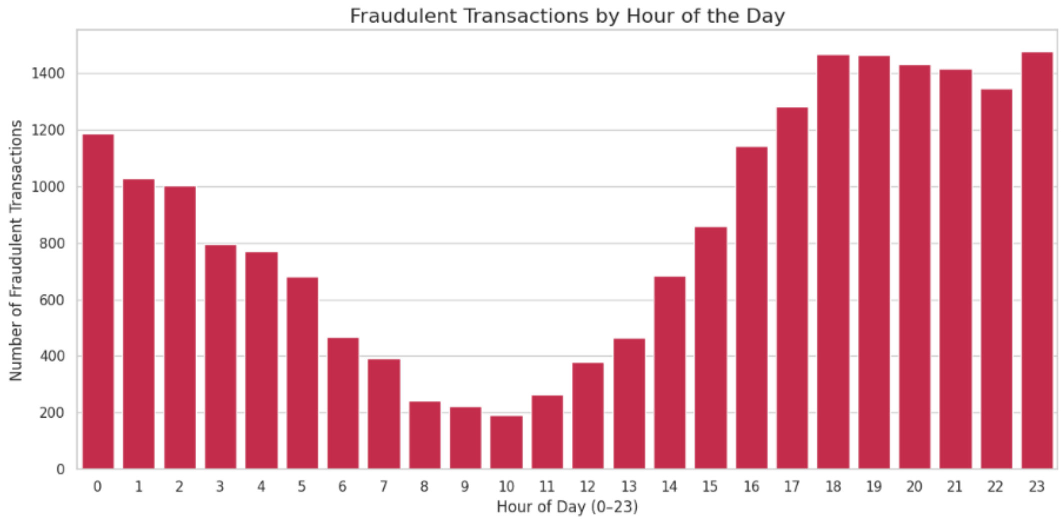
Benefits: 90% faster training time, dramatically improved recall.

Trade-off & Mitigation: Discarded ~528,000 legitimate transactions. Mitigated by rotating fresh samples during monthly retraining.



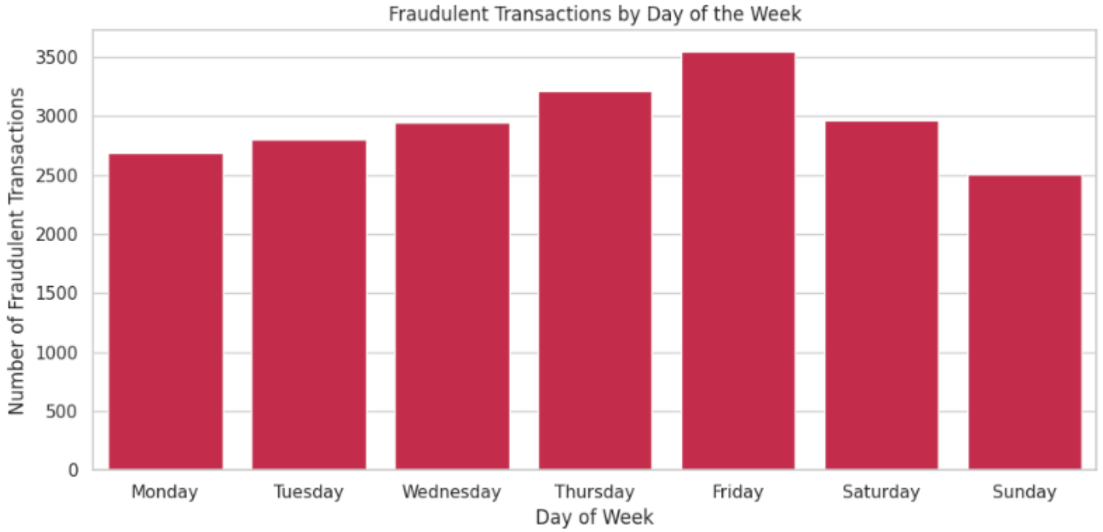
Fraud Patterns - Temporal Insights & Class Distribution

- **Source:** Exploratory Data Analysis (EDA) on the full dataset.
- **Class Imbalance:**
  - Fraudulent cases: 20,663 (3.5%)
  - Legitimate cases: 569,877 (96.5%)



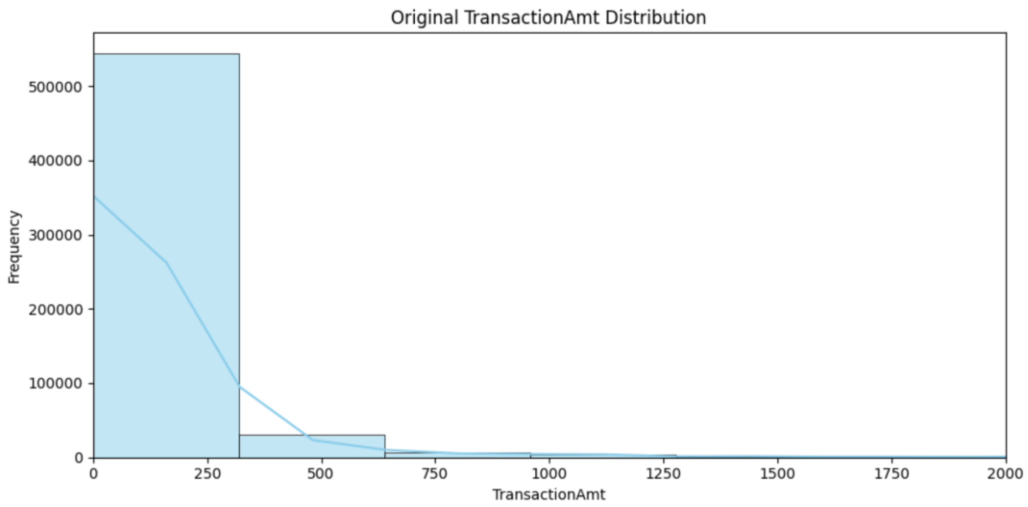
Key Fraud Patterns - Temporal Insights & Class Distribution

- **Hourly Distribution:**
  - Fraudulent activity spikes during **evening hours (4 PM to 11 PM)**.
  - Peak fraud at 11 PM (>1,300 transactions/hour).
  - Significant activity (20-25%) between midnight and 4 AM, indicating exploitation of reduced monitoring.



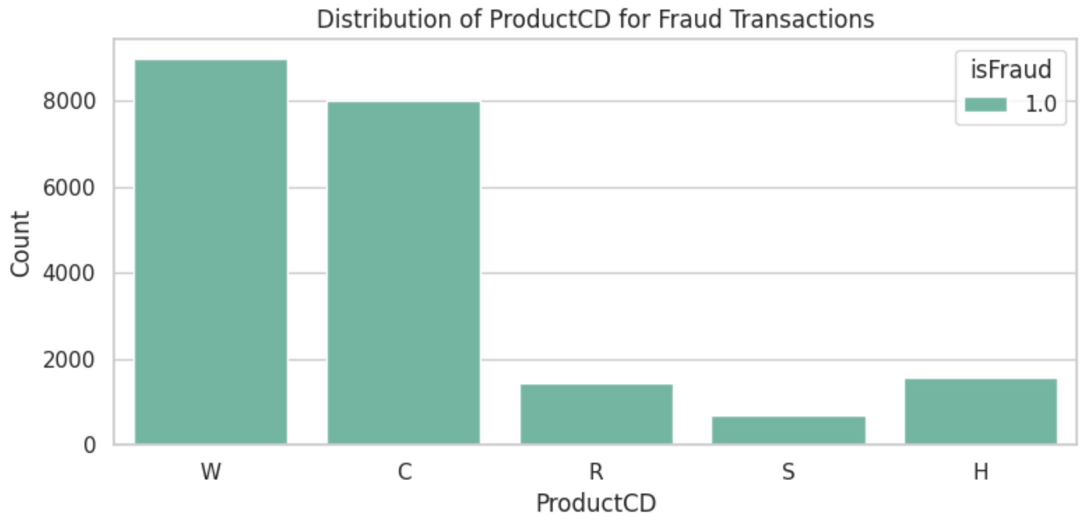
Key Fraud Patterns - Temporal Insights & Class Distribution

- **Weekday Distribution:**
  - **Fridays** account for nearly 18% of fraudulent activity (disproportionate to total transactions).
  - Suggests opportunistic behavior exploiting end-of-week vulnerabilities.



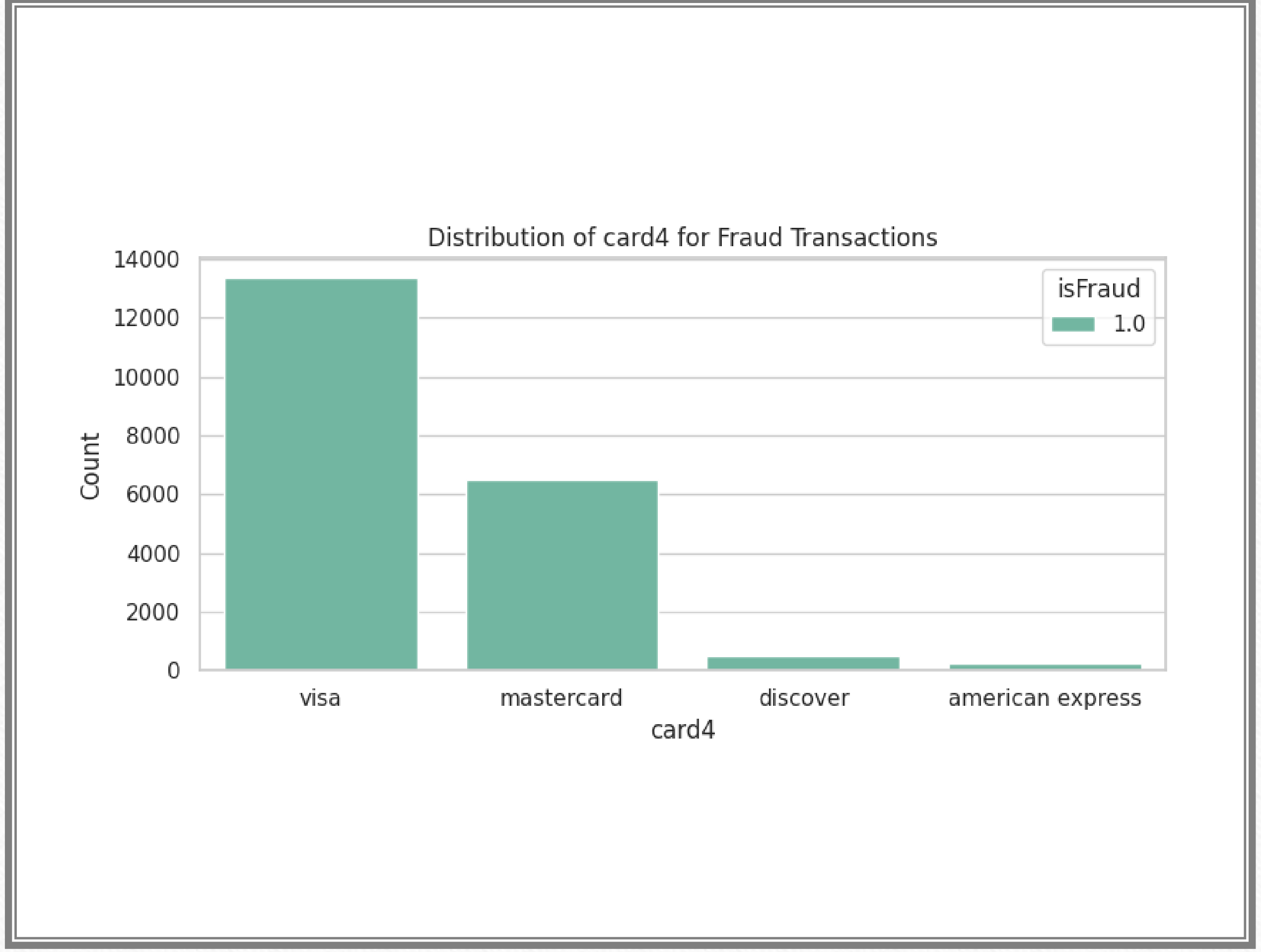
Key Fraud Patterns - Transaction & Network Characteristics

Transaction Amount: Most fraudulent transactions cluster between **$0 and $200**. Suggests small "test" charges.



Key Fraud Patterns - Transaction & Network Characteristics

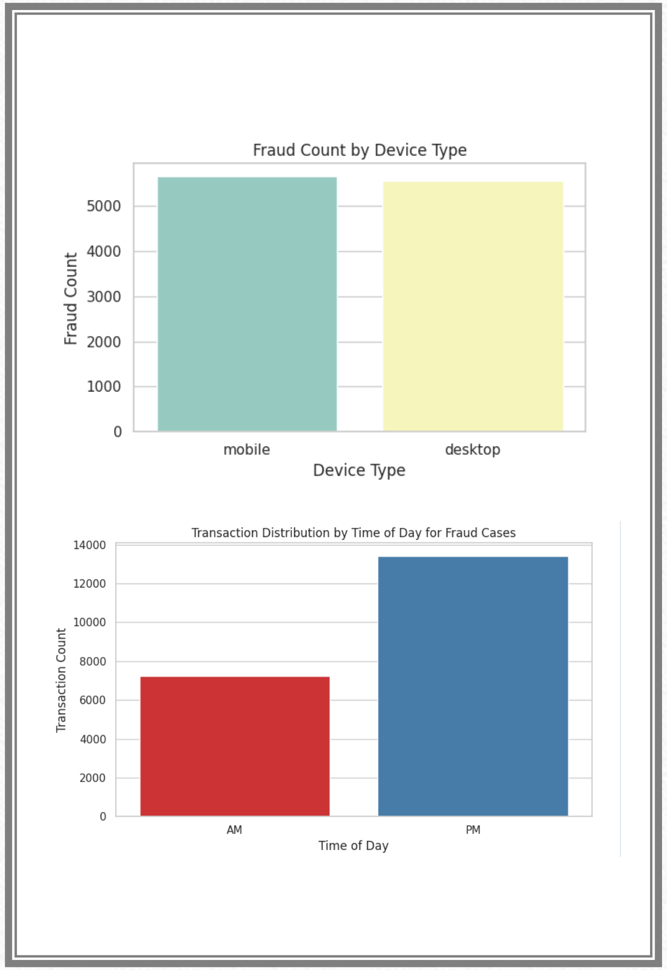
Product Code (ProductCD): Codes "W" and "C" are overrepresented in fraud cases (e.g., "W" is 5% of transactions but 12% of fraud).



Key Fraud Patterns - Transaction & Network Characteristics

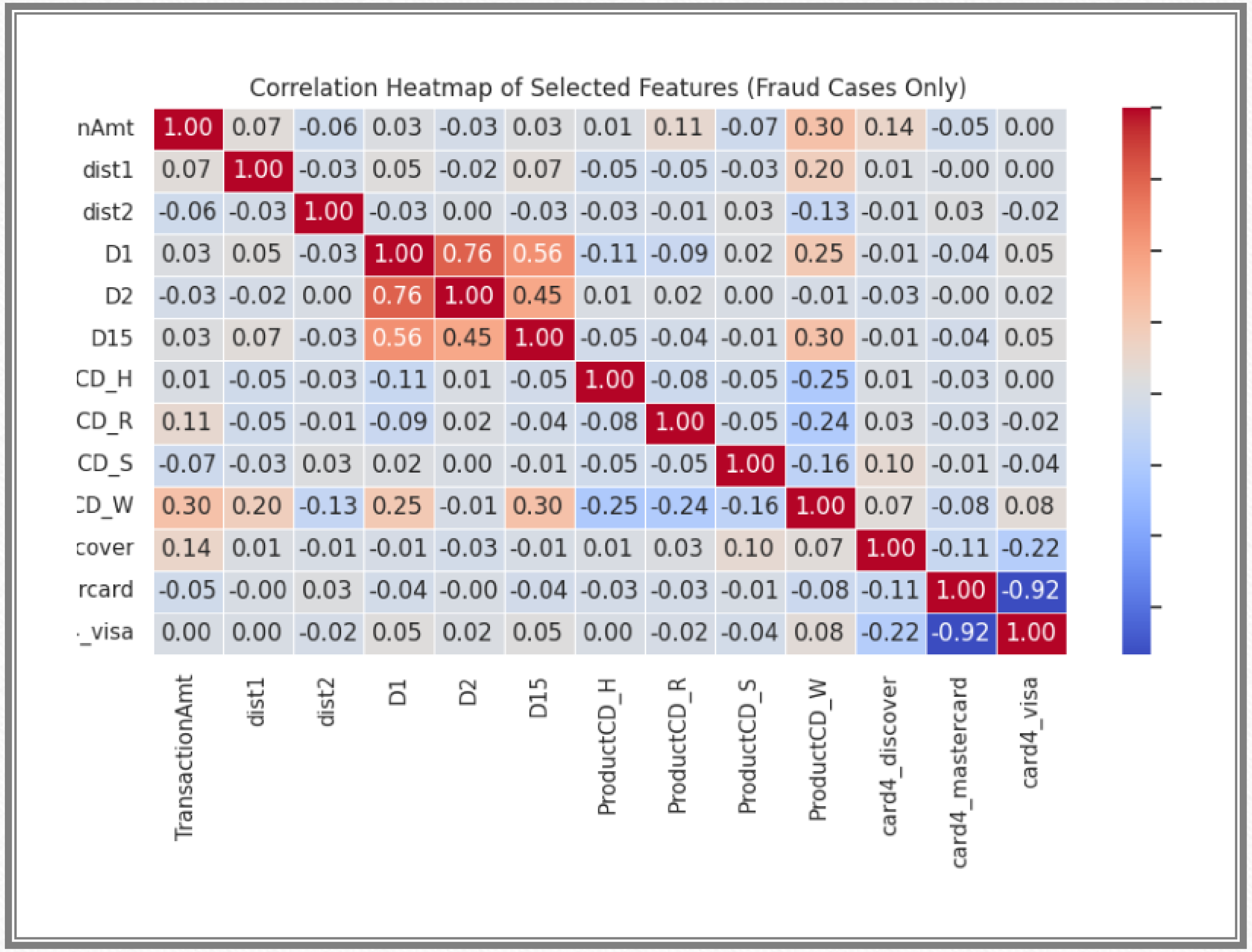
- **Card Network (card4):**
  - **Discover & MasterCard** show disproportionately higher fraud rates.
  - Example: Discover processes 18% of transactions but accounts for 28% of fraud.

Key Fraud Patterns - Device, and Behavioral Signals



Device Type: Nearly 55% of fraud originates from **mobile devices**.

Time of Day (AM vs. PM): PM period (afternoon/evening) sees ~1.4x more fraud than AM.



Key Fraud Patterns - Device, and Behavioral Signals

- **Distance & Time-Difference Correlations:**
  - Strong correlation (0.76) between D1 and D2 suggests rapid successive transactions.
  - Moderate correlation (0.30) between TransactionAmt and D15 implies higher fraud amounts follow previous transactions quickly.

Model Selection - Overview of Classifiers

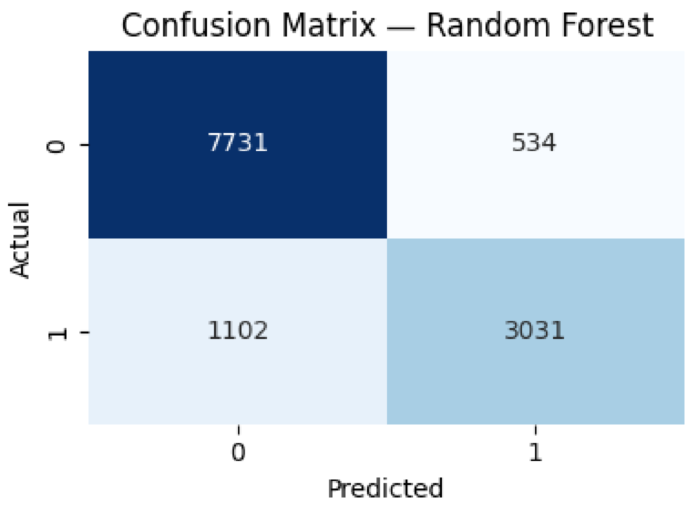
Goal: Build a robust fraud detection system balancing performance, efficiency, and scalability.

Classifiers Evaluated:

- **Logistic Regression:** Transparent baseline, high recall (99.27%) but impractical due to ~99% false positives.
- **Random Forest:** Ensemble of decision trees.
- **XGBoost:** Gradient-boosted decision trees.
- **LightGBM:** Optimized gradient-boosting framework.

Evaluation Focus: Prioritized high recall, balanced with acceptable precision, F1 score, and ROC AUC.

Training Data: All models trained on SMOTE-balanced training set (66,122 rows, 50/50 fraud/non-fraud).



Random Forest - The Chosen Model's Performance

- **Metrics (Test Set, default 0.50 threshold):**
  - **Accuracy:** 86.80%
  - **Precision (Fraud):** 85.02%
  - **Recall (Fraud):** 73.34%
  - **F1 Score (Fraud):** 78.75%
  - **ROC AUC:** 0.9176
- **Impact:**
  - Correctly detected 3,031 of 4,133 fraud cases (1,102 false negatives).
  - Misclassified only 534 of 8,265 legitimate transactions as fraud (false positives).
- **Why Chosen:** Offers the optimal balance between detecting fraud (recall) and minimizing false alarms (precision), strong class separation, and interpretability through feature importance.

Performance Comparison Across Models

Model	ROC AUC	Accuracy (%)	Precision(Fraud)(%)	Recall (Fraud)(%)	F1 Score (Fraud)(%)	False Positives	False Negatives
Logistic Regression	0.4990	33.45	33.29	99.27	49.86	8,182	28
Random Forest	0.9176	86.80	85.02	73.34	78.75	534	1,102
XGBoost	0.9146	86.44	85.13	71.37	77.64	501	1,183
LightGBM	0.9141	86.30	84.94	70.95	76.88	478	1,200

- **Random Forest:** Strong balance of recall (73.34%) and precision (85.02%), highest ROC AUC (0.9176).
- **XGBoost:** Very competitive, slightly lower recall (71.38%) but similar precision (85.48%), strong ROC AUC (0.9137).
- **LightGBM:** Fastest training, lowest false positives (478), but slightly lower recall (70.97%).
- **Logistic Regression:** Unacceptable false positives despite high recall.

Crucial Impact of Threshold Optimization

Default 0.50 Threshold: For Random Forest, this would lead to ~6,240 false positives daily (assuming 100k daily transactions), overwhelming manual review.

Optimized Threshold: 0.69:

- Crucial for operational feasibility.
- Maintains ~73% fraud recall.
- Reduces false positives to **under 1,000 daily**.

Business Impact: Projected annual savings of over $104,000 by balancing missed fraud costs and false alert overhead.

Recommendation 1: Real-Time Deployment

Model: Random Forest (300 estimators, max_depth=15, min_samples_split=5).

Deployment: Integrated into real-time transaction pipeline for immediate scoring.

Flagging Criteria: Transactions with fraud probability ≥ 0.69 flagged for manual review.

Monitoring: Daily precision, recall, and false positive counts. Log key features for flagged cases.

Expected Outcome: ~73% fraud detection with fewer than 1,000 false positives daily, making manual review sustainable.

Recommendation 2: Three-Tier Review Workflow

TIER	PROBABILITY RANGE	ACTIONS
1	$0.69 \le p < 0.80$	Queue for hourly batch review; send automated email alerts to the fraud operations team.
2	$0.80 \le p < 0.90$	Temporarily hold transactions; trigger real-time phone/SMS One-Time Password (OTP) verification for user authentication.
3	$p \ge 0.90$	Immediately suspend/decline the transaction; escalate the case for overnight manual investigation.

- **Purpose:** Optimize resource allocation based on predicted fraud probability.
- **Benefit:** Real-time intervention for highest-risk cases, efficient batch processing for moderate-risk.

Recommendation 3: Continuous Monitoring & Adaptive Retraining

Monthly Retraining Pipeline:

- Collect 60 days of recent labeled transactions.
- Create new 2:1 undersampled subsets.
- Apply preprocessing and SMOTE.
- Retrain Random Forest, validate on holdout sets.
- Recalibrate thresholds; deploy if performance meets/exceeds standards.

Real-time Monitoring Dashboard:

- Track 7-day/30-day rolling averages of key metrics.
- Monitor feature distributions for drift (e.g., >10%).
- Alert on performance degradation (>2 percentage points).
- Ensure review team capacity matches flagged volumes.

Goal: Maintain model effectiveness against evolving fraud tactics and preserve operational efficiency.

Current Limitations

1. **Information Loss from Undersampling:** Discarded ~528,000 legitimate transactions; may obscure rare valid patterns, potentially increasing false positives in production.
2. **SMOTE Bias:** Synthetic oversampling may not reflect emerging fraud schemes.
3. **Feature Drift Vulnerability:** Fraud patterns evolve, requiring vigilant monitoring to prevent performance degradation.
4. **Limited Network Analysis:** Current approach examines individual transactions; misses coordinated fraud rings.
5. **Reduced Interpretability for "V" Features:** Anonymized features limit understanding of model decisions.

Future Research Directions

- **Ensemble Methods:** Combine Random Forest, XGBoost, LightGBM for enhanced performance and robustness.
- **Graph-Based Anomaly Detection:** Leverage transaction-entity networks to identify coordinated fraud rings.
- **Streaming Learning Implementation:** Enable near-real-time model updates for faster adaptation to evolving tactics.
- **Behavioral Biometrics Incorporation:** Add user interaction patterns (mouse movements, typing dynamics) for richer fraud scoring.
- **Cost-Sensitive Optimization:** Develop models that directly optimize business impact, balancing false negative costs against false positive burdens.
- **Advanced Feature Engineering:** Explore temporal sequence modeling, peer-group comparisons, velocity-based features.

Conclusion - A Robust & Evolving Defense

Objective Achieved: Balanced high recall with acceptable precision in credit card fraud detection.

Key Achievements:

- Optimized Random Forest performance (73.34% recall, 85.02% precision, 0.9176 ROC AUC).
- Enhanced operational efficiency through 0.69 threshold (reduces false positives to <1,000 daily).
- Critical pattern discovery (temporal spikes, low amounts, mobile device risk).
- Established a scalable framework (monthly retraining, continuous monitoring).

Strategic Impact: Provides financial institutions with a robust, evolving defense against sophisticated fraud, safeguarding assets and customer trust.

Questions & Discussion

Thank you!