---

Create a 10-slide professional PowerPoint presentation titled

Detecting Fraud in Real Time: Sequence-Based Feature Engineering and Modeling in Telecom

This presentation is for a Data Science Summit (15-minute talk).

Keep the title exactly the same (it is already submitted in the abstract).

The deck should visually explain how I enhanced an existing 15-feature tabular fraud detection model by adding sequence-based features generated from Intelligence (MemoNodes) data, where each record represents a Band Number joined with the Billing Account Number.

---

Structure and Content (Slide by Slide)

Slide 1 – Title and Objective

Title exactly as above.

Goal: Integrate new sequence features on top of existing 15 tabular features to improve real-time fraud detection.

Mention that the data comes from the Intelligence data source and contains per-Band Number MemoNodes logs.

Slide 2 – Problem Context

Telecom fraud happens as short, rapid sequences of memo-type activities within 48 hours.

Static or tabular models miss temporal dependencies.

Objective: detect evolving behavior patterns across time windows.

Slide 3 – Data Overview

Item | Detail

Data Source | Intelligence (MemoNodes)

Primary Entity | Band Number (merged via Billing Account Number)

Scale | Around 2.5 lakh records, 850 frauds (0.3 percent)

Split | 200 K train / 50 K test (250 frauds in test)

Class Weight | 1 : 500

Windowing | Rolling 48-hour event sequences

Slide 4 – Data Preparation Workflow

Extract logs from Intelligence → MemoNodes.

Group by Band Number.

Remove duplicate memos at identical timestamps.

Order chronologically.

Apply chi-square significance test to reduce 458 memo types to 15 key memo types.

Merge with Billing Account Number attributes.

Slide 5 – Base vs Enhanced Model

Comparison Table:

Base Model | Enhanced Model (Our Approach)

15 tabular features | 15 + sequence-based features

GBM (class-weighted 1:500) | GBM (same algorithm)

Point-in-time analysis | 48-hour temporal sequence analysis

Goal: Fraud classification | Goal: Temporal behavioral fraud detection

Slide 6 – Sequence Feature Set Overview

Grouped feature categories:

Statistical Time: sequence length, min/max/median gap, standard deviation of time gaps.

Rarity: Average Rarity (IDF) per sequence.

Transition: bigram probabilities, sequence probability (product of conditional probabilities), entropy.

Stickiness/Streak: maximum consecutive same-type runs, repeat ratio.

Density and Entropy: events per hour, unique memo ratio.

Deep Encoding: Autoencoder (Bi-LSTM) reconstruction error score.

Slide 7 – Feature Transformation Logic (How Each Feature Was Generated)

Table format with three columns: Step, Transformation Logic, Example or Formula.

1. Form bigram tokens using underscores, e.g., ELG\_1100, 1100\_CER.

2. Compute time gaps between current and previous memo times (within 48 hours).

Formula: Δt = current\_memo\_time − previous\_memo\_time.

3. Derive statistical time features such as min, median, max, and standard deviation of time gaps.

4. Build a transition matrix and normalize rows. Formula: P(j|i) = freq(i,j) / Σfreq(i,\*).

5. Compute sequence probability as the product of all transition probabilities P(next|current).

6. Compute stickiness or streak features such as maximum repeated memo length and repetition ratio.

7. Pass sequence to Bi-LSTM Autoencoder and use reconstruction error as an additional anomaly score feature.

Slide 8 – Average Rarity (IDF-Based Feature)

Step-by-step table:

1. Collect all sequences per Band Number (examples: “A,B,C”, “B,D”, “A,C,D,E”).

2. Count how many sequences each memo token appears in.

3. Compute rarity using the formula rarity(t) = log₂(N / (1 + df(t))).

4. Broadcast the rarity map across Spark nodes.

5. Compute average token rarity per sequence as the mean of all token rarities.

6. Add this numeric feature as avg\_token\_rarity.

Example: If N = 4 and token A appears in 3, log₂(4/3) = 0.415; token B appears in 4, log₂(4/4) = 0.

Average rarity = mean of those values.

Slide 9 – Model Performance and Results

Comparison table with metrics:

Metric | Base Model (15) | Enhanced Model (15 + Sequence) | Gain

AUC | 0.80 | 0.90 | +12 percent

AUC-PR | 0.005 | 0.010 | 2x improvement

F1 | 0.67 | 0.78 | +16 percent

Detection latency reduced, with fraud flagged several hours earlier.

Slide 10 – Summary and Future Scope

Combined base tabular signals with sequence intelligence from Intelligence (MemoNodes).

Key new features: Average Rarity, Sequence Probability, Stickiness, Standard Deviation of Time Gap, Autoencoder Error, Density, and Entropy.

Significant improvements in AUC, AUC-PR, and F1.

PySpark pipeline supports real-time scoring and feature refresh.

Next steps: deploy streaming alerts and extend framework to finance and cybersecurity use cases.

---

Design Guidance

Use a professional blue-white theme similar to AT&T branding.

Include one architecture workflow diagram showing the flow:

Intelligence → 48-hour window and deduplication → underscore bigrams and time gaps → feature extraction → union with base 15 → weighted GBM → fraud probability score.

Highlight key new features: Average Rarity, Sequence Probability, Stickiness, Standard Deviation of Time Gaps, and Autoencoder Score.

Add light animations or arrows to show feature transformation flow.