Excellent 👏 — thank you for the clarity that Billing Account Number (BAN) is your unique entity ID, and that you’ll now have 20 minutes for the presentation.

That extra time lets you give a stronger narrative flow — a little slower pacing, slightly more visuals, and 1–2 more slides for clarity on data preparation and feature generation (especially statistical and probabilistic parts).

Below is your final structured version — optimized for a 20-minute Data Science Summit presentation.

I’ll outline each slide title, content, and narration plan (what to speak).

After this, if you confirm, I’ll generate the ready-to-use PowerPoint deck (.pptx) with all slides formatted.

---

🎯 Data Science Summit 2025 – Final 20-Minute Deck

“Sequence-Driven Feature Engineering for Enhanced Anomaly Detection”

R. Bhadra Rao | Specialist Data Scientist

---

🟩 Slide 1 — Title & Objective

Sequence-Driven Feature Engineering for Enhanced Anomaly Detection

R. Bhadra Rao | Specialist Data Scientist

Goal

Transform sequential customer or system activity logs into explainable numerical features that enhance anomaly/fraud detection accuracy.

Highlights

48-hour sequence extraction per Billing Account Number (BAN)

Statistical • Rarity • Probabilistic • Autoencoder features

Integrated with LightGBM baseline → +20 % AUC, ×2 AACPR

🗣 Speak:

“Today I’ll present how we engineered intelligence from raw sequential data — converting event logs into predictive features that doubled our detection power.”

---

🟩 Slide 2 — Business Context & Motivation

Background

Each BAN represents a unique customer or entity

Event logs capture multiple activities for every BAN

The last event = application processing time (ground truth)

Goal: model patterns within 48 hours preceding that point

🗣

“This work helps identify unusual behavioral sequences before a key event. Instead of just static attributes, we analyze temporal patterns.”

---

🟩 Slide 3 — Data Architecture Overview

Inputs

Base dataset → 15 tabular features

Log dataset (Intelligence source) → ~4 lakh records, ~450 memo types

Join Logic

Key Field Description

Billing Account Number (BAN) Unique identifier

Application ID Process instance ID

Processing Time Ground truth timestamp

Merge Rule → Base Table ⨝ Logs on (BAN, Application ID)

→ Select events within 48 hours before Processing Time

🗣

“This merge provides a complete 48-hour activity window for each BAN–application pair.”

---

🟩 Slide 4 — Pre-Processing Steps

1. Filter 48-hour logs per BAN + App ID

2. Remove duplicate memo types

3. Sort events chronologically (oldest→latest)

4. Compute time\_gap = processing\_time − event\_time (in seconds)

5. Generate ordered sequence memo\_type\_ordered

6. Store time\_gap\_array and order\_length

BAN App ID Processing Time Event Seq time\_gap\_array (s) order\_len

BAN\_101 APP\_001 10:00 A\_B\_C [18000, 7200, 1800] 3

BAN\_102 APP\_002 12:00 B\_D\_E [36000, 10800, 1800] 3

🗣

“This transformation standardizes all sequences for analysis, aligning by seconds for precision.”

---

🟩 Slide 5 — Statistical Features from Time Gaps

BAN Sequence min max median stddev order\_len

BAN\_101 A\_B\_C 1800 18000 7200 6710 3

BAN\_102 B\_D\_E 1800 36000 10800 15100 3

BAN\_103 D\_E\_C\_E 1800 86400 36000 31800 4

🧩 Features Generated

min\_time\_gap, max\_time\_gap, median\_time\_gap, stddev\_time\_gap, order\_length

🗣

“These metrics reflect the spread and intensity of activities within the 48-hour window.”

---

🟩 Slide 6 — Average Rarity (IDF-Based Feature)

Process

Step Action Example

1 Count how many BANs contain each memo type (df) A: 3 B: 3 C: 2 D: 2 E: 3 F: 1

2 Compute rarity = log₂(N / (1+df)) A=0.32 F=1.00

3 Replace each token with rarity value A\_B\_C → [0.32, 0.32, 0.42]

4 Take average → avg\_token\_rarity 0.35

🗣

“This is similar to IDF in text analytics — sequences containing rare event types receive higher rarity scores.”

---

🟩 Slide 7 — Contingency Frequency Table

Raw Transition Counts (from all BANs)

From → To A B C D E F Row Total

A – 20 40 0 10 0 70

B 0 – 35 15 0 0 50

C 10 0 – 0 25 0 35

D 0 0 0 – 30 0 30

E 5 0 10 0 – 0 15

🗣

“This shows how often each event is followed by another — it’s the foundation for our probabilistic transition modeling.”

---

🟩 Slide 8 — Probability Contingency Table (Normalized)

| From → To | A | B | C | D | E | Σ P(next | current) | |:--|--:|--:|--:|--:|--:|--:| | A | – | 0.29 | 0.57 | 0 | 0.14 | 1.00 | | B | 0 | – | 0.70 | 0.30 | 0 | 1.00 | | C | 0.29 | 0 | – | 0 | 0.71 | 1.00 | | D | 0 | 0 | 0 | – | 1.00 | 1.00 | | E | 0.33 | 0 | 0.67 | 0 | – | 1.00 |

Derived Features

prob\_product = Π P(next|current)

prob\_entropy = –Σ P log₂ P

prob\_geo\_mean = (Π P)^(1/n)

Example

BAN Sequence Pairs prob\_product prob\_entropy geo\_mean

BAN\_101 A\_B\_C (A→B),(B→C) 0.29×0.70=0.20 1.22 0.45

BAN\_103 B\_D\_E\_C\_E … 0.30×1×0.67×0.71=0.14 1.68 0.45

🗣

“These features represent the probability consistency and uncertainty of each BAN’s sequence transitions.”

---

🟩 Slide 9 — Bi-LSTM Autoencoder Feature

BAN Input Sequence Reconstructed Sequence AE\_Error

BAN\_101 A\_B\_C\_D\_E A\_B\_C\_D\_E 0.18

BAN\_103 B\_D\_E\_C\_E B\_D\_X\_C\_E 0.52

BAN\_104 A\_F A\_X 0.71

🧠 Concept

Trains on normal patterns using Bi-LSTM Autoencoder

Reconstruction error → degree of anomaly

🗣

“Sequences with higher reconstruction error are those that deviate from normal patterns — a strong anomaly indicator.”

---

🟩 Slide 10 — Consolidated Feature Groups

Group Features Purpose

Statistical (Time Gaps) min, max, median, stddev Temporal spread

Sequence Structure order\_length Activity density

Rarity (IDF) avg\_token\_rarity Behavioral uniqueness

Probabilistic (Markov) prob\_product, entropy, geo\_mean Transition likelihood

Autoencoder ae\_error Anomaly score

🗣

“This complete feature set captures both the temporal and structural behavior of each BAN.”

---

🟩 Slide 11 — Model Integration

Model Features Description

Base LightGBM 15 tabular features Existing baseline

Enhanced LightGBM 15 + sequence features Added statistical, rarity, probabilistic & AE signals

Training Setup

2.5 L records (~850 rare cases)

Split 80 : 20 • Class weight 1 : 500

Metrics → AUC, AACPR, F1

🗣

“No change to the algorithm — just smarter features.”

---

🟩 Slide 12 — Model Performance Comparison

Metric Base Model Enhanced Model Improvement

AUC 0.801 0.956 +19 %

AACPR 0.050 0.100 +100 %

F1 Score 0.48 0.69 +43 %

🗣

“The enhanced features significantly improved both discrimination and recall, even on imbalanced data.”

---

🟩 Slide 13 — Insights & Observations

✅ Sequence-aware features add critical context

✅ Rarity + transition probabilities reveal hidden signals

✅ Autoencoder detects structural anomalies

✅ Feature engineering alone yielded 20 % AUC gain

🗣

“This shows how powerful well-designed features can be — we achieved major improvements without complex model changes.”

---

🟩 Slide 14 — End-to-End Workflow

Intelligence Logs

↓

Pre-Processing (48 h window per BAN)

↓

Sequence Generation (memo\_type\_ordered, time\_gap\_array)

↓

Feature Derivation

├─ Statistical Features

├─ Average Rarity (IDF)

├─ Probabilistic (Transition Matrix)

└─ Autoencoder Anomaly Score

↓

Enhanced LightGBM Model

↓

AUC 0.801 → 0.956 AACPR 0.05 → 0.10

🗣

“This pipeline converts sequential logs into rich, explainable features driving measurable improvement.”

---

🟩 Slide 15 — Impact & Takeaways

📈 AUC ↑ 19 % AACPR × 2

⚙️ Fully scalable and interpretable features

🧠 Combines statistical, probabilistic, deep sequence signals

💡 Feature innovation delivers real business value

> “We transformed sequential activity data into predictive intelligence — proving that feature design can redefine model performance.”

---

🟩 Slide 16 — Q&A + Acknowledgment

Thank you!

Questions & Discussion

> “This was a collaborative journey of engineering, analytics, and domain understanding — glad to share it here.”

---

⏱ 20-Minute Delivery Flow

Section Slides Time (min)

Intro & Context 1–2 2

Data Prep 3–4 3

Statistical & Rarity 5–6 3

Probabilistic 7–8 4

Autoencoder & Feature Summary 9–10 3

Model Integration & Results 11–12 3

Insights & Closing 13–16 2

---

Would you like me to now generate this as a PowerPoint file (.pptx) with clean formatting, consistent font, modern color theme (white background + accent blue/green), and ready for presentation upload?