Novartis Datathon 2025: Complete Comprehensive Official Guide

Generic Erosion Forecasting Challenge

Barcelona Digital Finance Hub

Table of Contents

---

1. Executive Summary

The **Novartis Datathon 2025** is a data science competition hosted by the **Barcelona Digital Finance Hub**. Participants are challenged to **forecast pharmaceutical sales volume erosion** following the entry of generic competitors into the market.

The Problem

When a drug's patent expires (Loss of Exclusivity), generic manufacturers can legally produce and sell equivalent versions at lower prices. This typically causes significant declines in the originator brand's sales—a phenomenon known as **generic erosion**. Accurately forecasting this erosion is critical for:

- Revenue forecasting
- Production planning
- Strategic decision-making
- Portfolio management

Key Objectives

Objective	Description
Primary Goal	Predict monthly sales volumes for 24 months following generic entry
Focus Area	High-erosion drugs (Bucket 1) that lose ≥75% of pre-generic sales
Business Context	Support financial planning and strategic decisions during the post-patent period

Two Forecasting Scenarios

Scenario	Timing	Forecast Horizon	Available Data
Scenario 1	Immediately at generic entry	Months 0–23 (24 months)	Pre-generic history only
Scenario 2	Six months after entry	Months 6–23 (18 months)	Pre-generic + 6 months post-entry

Competition Structure

Phase	Description	Teams Evaluated	Advancement
Phase 1-a	Scenario 1 accuracy	All teams	Top 10 advance
Phase 1-b	Scenario 2 accuracy	Top 10	Top 5 advance
Phase 2	Jury presentation	Top 5	**Top 3 winners selected**

Dataset Summary

Dataset	Observations	Purpose
Training Set	1,953 country-brand pairs	Model development
Test Set - Scenario 1	228 observations	Forecast Months 0-23
Test Set - Scenario 2	112 observations	Forecast Months 6-23
Total	**2,293 combinations**	

---

2. Introduction and Context

2.1 About the Datathon

The Novartis Datathon 2025 is hosted by the **Barcelona Digital Finance Hub**, a center dedicated to applying data science and advanced analytics to financial processes within Novartis. This challenge brings together data enthusiasts, innovators, and problem solvers from around the world to tackle a real-world, high-impact problem at the intersection of:

- Pharmaceutical business strategy
- Financial planning
- Advanced analytics and forecasting

2.2 Novartis Mission

At the core of this initiative stands **Novartis' mission**:

> "*To reimagine medicine in order to improve people's lives.*"

Novartis is committed to:

- Using **innovative science and technology** to address healthcare challenges
- Discovering and developing **breakthrough treatments**
- Finding new ways to deliver treatments to as many patients as possible

2.3 The Central Problem

The Datathon focuses on **generic erosion**—the sharp decline in sales volume that branded drugs experience after generic competitors enter the market following patent expiry. Accurate prediction of this erosion is critical for:

- Revenue forecasting
- Production planning
- Strategic decision-making

- Managing the post-patent period

---

3. The Barcelona Digital Finance Hub

3.1 History and Growth

Year	Team Size
2018	10
...	...
2025	55

The hub has grown more than **fivefold** from 2018 to 2025.

3.2 Team Composition (2025)

- **34 Data Scientists** – Build statistical and machine-learning models, create predictive analytics
- **8 Finance Professionals** – Translate business questions into data problems, validate results
- **15 Engineers** – Visualization specialists, ML engineers, Software engineers, DevOps

3.3 Team Diversity

- **14+ Different nationalities** represented
- **66% Local talent** from Barcelona/Spain
- **34% International** talent

3.4 Educational Background

Discipline	Percentage
Mathematics & Statistics	24%
Computer Science	20%
Economics	20%
Physics & Others	19%
Engineering	17%

Additional qualifications:

- **5 PhD holders**
- **3 Bioinformatics specialists**

3.5 Why Barcelona?

Barcelona was chosen as a strategic location due to:

1. **Tech Cluster**: Amazon, Microsoft, AstraZeneca, and other companies have located their global AI hubs in Barcelona
2. **Research Infrastructure**: Barcelona Supercomputing Center, Quantum Computer, Synchrotron
3. **Academic Excellence**: Strong university programs in Data Science, Mathematics, and Statistics (UPF, UPC, UB)
4. **Talent Attraction**: Quality of life, climate, culture, and cost of living make it attractive for international talent

3.6 Hub Mission

The Digital Finance Hub serves as a **bridge between finance and digital innovation**, enabling:

- Data-driven decision-making at scale
- Application of big data and AI to financial processes
- Transformation of pharmaceutical industry finance operations

---

4. Pharmaceutical Background

4.1 Patents and Loss of Exclusivity (LOE)

When a pharmaceutical company develops a new drug, it receives a **patent** granting exclusive rights to produce and commercialize the product for a limited period.

Patent Protection

Aspect	Details
Duration	Typically 20 years from filing
Rights Granted	Exclusive manufacturing and commercialization
Purpose	Allow company to recoup R&D investments
Scope	No other company can produce the same drug without permission

Loss of Exclusivity (LOE)

Loss of Exclusivity (LOE) occurs when:

- The patent expires
- Legal protection ends
- Generic manufacturers can legally enter the market

Timeline Progression:

```
Innovation → Patent Grant → Exclusivity Period → LOE → Open Competition
```

4.2 Generic Drug Definition

A **generic drug** must be therapeutically equivalent to the brand-name (originator) medication.

Equivalence Requirements

Attribute	Description
Dosage Form	Tablet, capsule, injectable, etc.
Strength	Amount of active ingredient per dose (e.g., 80 mg)
Route of Administration	Oral, intravenous, topical, etc.
Quality	Meets regulatory standards for purity and stability
Performance	Behaves similarly in the body
Intended Use	Same indications and patient population

Important: Generic products may contain different **inactive ingredients** (fillers, binders, colorants, coatings), but these differences must not affect therapeutic outcomes. The active pharmaceutical ingredient must be identical.

4.3 Bioequivalence Requirements

Generic manufacturers do not repeat full clinical trials. Instead, they must demonstrate **bioequivalence** through pharmacokinetic studies.

Pharmacokinetic Properties Compared

Property	Definition
Absorption	Rate and extent of drug entering the bloodstream
Distribution	How the drug spreads through body tissues
Metabolism	How the body transforms (breaks down) the drug
Elimination	How the drug and metabolites are excreted

Typical Bioequivalence Study Protocol

1. **Study Design**: Healthy volunteers receive both brand-name and generic products (crossover design)
2. **Sample Collection**: Blood samples collected over time
3. **Analysis**: Concentration-time curves (pharmacokinetic profiles) analyzed
4. **Key Metrics**: AUC (area under curve) and Cmax (maximum concentration) compared
5. **Acceptance Criteria**: Products considered bioequivalent if ratios fall within **80-125% acceptance range**

This streamlined approval process significantly reduces development costs, enabling generics to be sold at lower prices.

4.4 Market Consequences of Generic Entry

Generic entry typically leads to significant market changes:

Consequence	Description	Impact
Increased Competition	Multiple manufacturers produce the same medication	Drives prices downward
Improved Affordability	Lower development costs enable lower prices	More accessible treatments
Greater Access	Reduced prices expand treatment availability	Better disease management outcomes
Substitution Practices	Pharmacists may substitute generics for branded products	Accelerates market shift
Revenue Decline	Originator loses market share rapidly	Financial impact on branded drug

4.5 Real-World Example: Diovan

Diovan (Novartis) illustrates the generics problem:

Attribute	Details
Active Ingredient	Valsartan (angiotensin II receptor blocker, ARB)
Indications	Hypertension (high blood pressure), heart failure
Patent Expiry	2012

Attribute	Details
Post-LOE Impact	Multiple generic manufacturers entered, significantly increasing competition and reducing prices
Result	Sharp decline in Novartis' Diovan revenue

This example illustrates why predicting and managing generic erosion is critical for pharmaceutical companies.

---

5. Drug Lifecycle and Generic Erosion

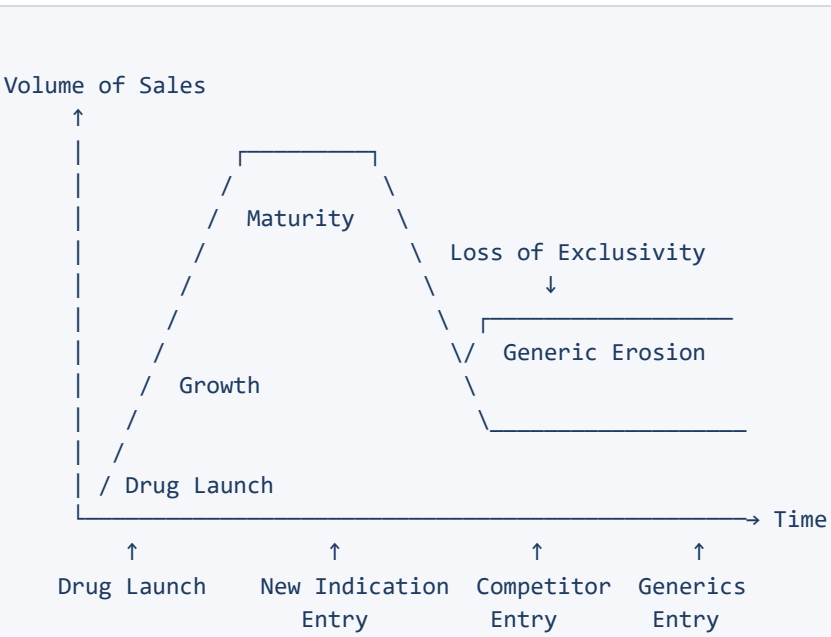
5.1 Complete Drug Lifecycle Phases

A pharmaceutical product progresses through distinct phases from market introduction to eventual sales decline:

```
Drug Launch → Growth → Maturity → Competition → Loss of Exclusivity → Generic Erosion
```

Phase	Characteristics	Sales Behavior
Drug Launch	Market introduction; low awareness and adoption	Sales low but rising
Growth	Increased prescriber awareness; market acceptance	Rapid sales increase
New Indication Entry	Additional disease/patient group approvals	Accelerated growth; upward curve
Maturity	Market saturated; established reimbursement	Sales stabilize at peak
New Competitor Entry	Other branded products enter same therapeutic area	Growth slows; curve flattens
Loss of Exclusivity (LoE)	Patent protection expires	Peak reached; decline begins
Generics Entry (Gx)	Generic competitors launch	Sharp, rapid volume decline
Generic Erosion	Post-generic stabilization	Low residual sales

Visual Representation



5.2 Key Commercial Milestones

Loss of Exclusivity (LoE)

Aspect	Details
Definition	Expiration of legal protections (patents, data exclusivity) preventing generic copying
Timing	Occurs near or at the peak of the sales curve
Impact	Allows generic manufacturers to launch cheaper copies

Generics Entry (Gx)

Aspect	Details
Definition	The moment generic drugs actually enter the market
Timing	Typically very close to, or just after, LoE
Impact	Triggers rapid decline in originator brand's sales volume

5.3 Generic Erosion Definition

Generic erosion is the steep and sudden decline in branded drug sales volume following generic entry. This is the **central topic** of the Datathon.

The Datathon Focus Window:

- Time window: From generic entry date through the subsequent **24 months**
- Focus: Forecasting the volume decline and understanding erosion dynamics
- **NOT** about the entire drug lifecycle—specifically about the post-LoE phase

Business Importance:

- Directly affects revenue forecasts
- Impacts production planning
- Influences strategic decisions (pricing, promotion, portfolio management)
- Enables preparation for post-patent period
- Helps minimize financial losses
- Supports competitive strategy adaptation

5.4 Mean Generic Erosion Metric

The **Mean Generic Erosion (MGE)** quantifies erosion severity over 24 months post-generic entry.

Formula 1: Mean Generic Erosion

$$\text{Mean Generic Erosion} = \frac{1}{24} \sum_{i=0}^{23} \text{Vol}_{\text{norm},i}$$

Where:

- i = month index (0 = entry month, 23 = 24th month after entry)
- $\text{Vol}_{\text{norm},i}$ = normalized volume in month i

Formula 2: Normalized Volume

$$\text{Vol}_{\text{norm},i} = \frac{\text{Vol}_i}{\text{Avg}_j}$$

Where:

- Vol_i = actual sales volume in month i after generic entry
- Avg_j = reference average volume for product/market j

Formula 3: Pre-Generic Reference Average (Baseline)

$$\text{Avg}*j = *frac{1}{12}* *sum*{i=-12}^{-1} Y^{\text{act}}_{j,i}$$

Where:

- j = drug/country/market index
- $Y^{\text{act}}_{j,i}$ = actual observed volume for drug/market j in month i
- $i = -12$ to -1 = the 12 months before generic entry

Interpretation of Normalized Volume

$\text{Vol}_{\text{norm},i}$ Value	Meaning
1.0	Current month equals pre-generic average
0.5	Volume is 50% of pre-generic average
0.1	Volume is 10% of pre-generic average
> 1.0	Volume exceeds pre-generic average (rare)

Interpretation of Mean Generic Erosion Values

MGE Value	Interpretation
≈ 1	Little erosion; post-generic volume similar to pre-generic
0.5	50% of pre-generic sales retained on average
≈ 0.25	Only 25% of pre-generic sales retained
≈ 0	Near-total sales collapse

5.5 Erosion Visualization Profiles

Three typical volume erosion patterns exist after generic entry:

High Erosion (Bucket 1 Profile)

- **Pre-entry:** Volume at stable high level
- **Post-entry:** Volume collapses almost vertically
- **Long-term:** Very low residual level near zero
- **Characteristics:** Steepest and deepest decline
- **MGE:** 0 to 0.25

Medium Erosion

- **Pre-entry:** Highest and stable volume
- **Post-entry:** Sharp fall but not to zero
- **Long-term:** Continues gradual decline over time
- **Characteristics:** Significant but not complete erosion
- **MGE:** 0.25 to 0.75 (approximately)

Low Erosion

- **Pre-entry:** Moderate, stable or slightly decreasing volume
- **Post-entry:** Slow, smooth decline
- **Long-term:** Remains relatively high among profiles
- **Characteristics:** Brand retains substantial market share
- **MGE:** 0.75 to 1.0

5.6 Erosion Classification and Buckets

Three Conceptual Categories

Category	Mean Erosion Range	Description
Low Erosion	Close to 1	Volume remains relatively stable
Medium Erosion	Between 0.25 and ~0.75	Moderate decline in sales
High Erosion	Close to 0	Sharp drop in volume

Two Datathon Buckets

For competition purposes, drugs are classified into **two operational buckets**:

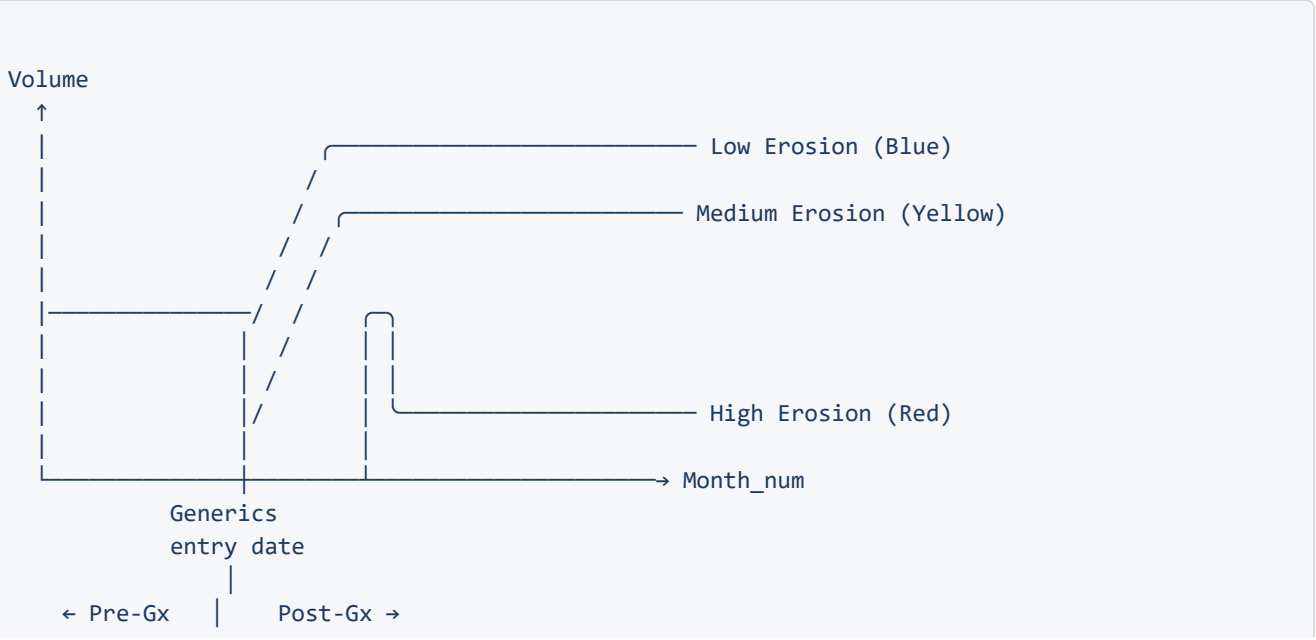
Bucket	Mean Erosion Range	Description	Focus Level	Metric Weight
Bucket 1 (B1)	[0, 0.25]	High erosion—loses ≥75% of pre-generic sales	**Primary Focus**	**2×**
Bucket 2 (B2)	(0.25, 1]	Medium/low erosion—retains >25% of pre-generic sales	Secondary	1×

Interval Notation:

- `[0, 0.25]` : Both 0 and 0.25 are **included** (square brackets)
- `(0.25, 1]` : 0.25 is **excluded** (round bracket), 1 is **included** (square bracket)

Important: Bucket labels are **NOT provided** in the datasets. You must derive them using the Mean Generic Erosion formula from the training data.

Visual Representation of Erosion Patterns



6. The Datathon Challenge

6.1 Core Objective

Forecast the **volume erosion** following generic entry over a **24-month horizon** from the generic entry date. This is a **time-series forecasting problem** focused on predicting monthly sales volumes for branded drugs after generic competitors enter the market.

6.2 Scenario 1: Zero-Knowledge Forecast

Aspect	Details
Timing	Immediately at generic entry date (month 0)
Available Data	Pre-generic history only; **no post-entry actuals**
Forecast Horizon	Month 0 to Month 23 (24 months)
Test Observations	228 country-brand pairs
Business Context	Planning at LOE before seeing market reaction
Challenge Level	Hardest—predict entire erosion trajectory without post-entry behavior

This scenario simulates: The situation where a country/brand has just lost patent protection and early planning is required without observed post-entry volumes.

6.3 Scenario 2: Six-Month Update

Aspect	Details
Timing	Six months after generic entry
Available Data	Pre-generic history + **6 months post-entry actuals** (months 0–5)
Forecast Horizon	Month 6 to Month 23 (18 months)
Test Observations	112 country-brand pairs
Business Context	Mid-course update with observed market data
Challenge Level	Moderate—use early erosion pattern to refine predictions

This scenario simulates: Updating and refining forecasts after observing the initial market reaction to generic entry.

6.4 Technical and Business Dimensions

The challenge is **not only** to build accurate models, but also to demonstrate **business understanding**.

Technical Requirements

- Build accurate forecasting models for both scenarios
- Handle data preprocessing appropriately
- Justify feature engineering decisions
- Document methodology clearly

Business Requirements

Requirement	Description
Deep Exploratory Analysis	Thoroughly explore the dataset: distributions, correlations, trends, differences across countries/molecules/therapy areas
Preprocessing Documentation	Explain and justify data cleaning choices: missing values handling, outlier treatment, normalization, feature engineering
High-Erosion Focus	Analyze characteristics of Bucket 1 markets: specific countries, therapeutic areas, competitive situations
Visualization Tools	Present time-series plots, comparative charts, feature importance, geographic/categorical breakdowns
Business Justification	Explain **why** approaches were chosen in terms of business value, interpretability, and practicality

7. Data Description

7.1 Dataset Overview

Data consists of historical monthly volumes for **2,293 country-brand combinations** that have experienced generic entry.

Data Structure: Three separate DataFrames

1. **Volume Dataset** (`df_volume.csv`) — Time-series of monthly sales
2. **Generics Dataset** (`df_generics.csv`) — Competition information over time
3. **Medicine Information Dataset** (`df_medicine_info.csv`) — Static product attributes

Join Keys: `country` and `brand_name`

7.2 Train-Test Split

Dataset	Observations	Data Availability
Training Set	1,953	Up to 24 months pre-entry + up to 24 months post-entry
Test Set - Scenario 1	228 (~67%)	Pre-entry data only
Test Set - Scenario 2	112 (~33%)	Pre-entry + 6 months post-entry
Total	**2,293**	

7.3 Volume Dataset (df_volume.csv)

The **core time-series dataset** containing monthly sales volumes around generic entry.

Schema

Column	Type	Description
`country`	String	Market identifier (anonymized, e.g., `COUNTRY_B6AE`)
`brand_name`	String	Brand identifier (anonymized, e.g., `BRAND_1C1E`)
`month`	String/Date	Calendar month of observation
`months_postgx`	Integer	Months relative to generic entry
`volume`	Float	**TARGET VARIABLE** — Number of units sold

`months_postgx` Interpretation

Value	Meaning
0	Month of generic entry
Negative (e.g., -3, -12)	Months before generic entry
Positive (e.g., 6, 23)	Months after generic entry

Example Data

country	brand_name	month	months_postgx	volume
COUNTRY_B6AE	BRAND_1C1E	Jul	-24	272594.39
COUNTRY_B6AE	BRAND_1C1E	Aug	-23	351859.31
COUNTRY_B6AE	BRAND_1C1E	Sep	-22	447953.48

Key Uses

- Reconstruct volume trajectory for each country-brand pair
- Identify pre-generic and post-generic periods
- Calculate baseline volumes for normalization
- Feed time-series and panel models for forecasting

7.4 Generics Dataset (df_generics.csv)

Contains **time-varying information on generic competitors** for each brand and country.

Schema

Column	Type	Description
`country`	String	Market identifier
`brand_name`	String	Brand identifier
`months_postgx`	Integer	Months after generic entry (starts at 0)
`n_gxs`	Integer/Float	Number of generic competitors at that time

Note: `n_gxs` varies over time as generics enter or exit the market.

Example Data

country	brand_name	months_postgx	n_gxs
COUNTRY_B6AE	BRAND_DF2E	0	0.0
COUNTRY_B6AE	BRAND_DF2E	1	0.0
COUNTRY_B6AE	BRAND_DF2E	2	1.0
COUNTRY_B6AE	BRAND_DF2E	3	2.0

Example Interpretation:

- Months 0-1: No active generic competitors (regulatory/commercial delays possible)
- Month 2: First generic appears

- Month 3+: Second generic joins, competition intensifies

Modeling Applications

- Join with `df_volume.csv` on `(country, brand_name, months_postgx)`
- Use `n_gxs` directly as a feature
- Derive additional variables:
  - Binary indicator: "any generics present" (`n_gxs > 0`)
  - Cumulative months since first generic appeared
  - Month-over-month change in `n_gxs`

7.5 Medicine Information Dataset (df_medicine_info.csv)

Contains **static product-level attributes** for each country-brand combination.

Schema

Column	Type	Description
`country`	String	Market identifier
`brand_name`	String	Brand identifier
`ther_area` / `therapeutic_area`	String	Therapeutic area (e.g., `Sensory_organs` , `Nervous_system`)
`hospital_rate`	Float	Percentage of drug delivered in hospitals (0-100)
`main_package`	String	Most common dispensing format (e.g., `PILL` , `INJECTION` , `EYE DROP`)
`biological`	Boolean	Whether drug is derived from living organism
`small_molecule`	Boolean	Whether drug is a low molecular weight compound

Column Details

`ther_area` (Therapeutic Area):

- Indicates clinical indication category
- Examples: `Sensory_organs` , `Muscoskeletal_Rheumatology` , `Antineoplastic` , `Nervous_system`
- Different areas show different erosion patterns due to clinical need, prescribing habits, and reimbursement rules

`hospital_rate` :

- Range: 0 to 100
- High values (e.g., 92%) indicate predominantly hospital-based distribution
- Low values (e.g., 0.09%) indicate mostly retail distribution
- High hospital share may indicate tender-driven procurement and stepwise erosion patterns

`main_package` :

- Categorical description of dosage/dispensing format
- Examples: `PILL` , `INJECTION` , `EYE DROP`
- Affects patient convenience, generic competition intensity, and price differentials

`biological` :

- `True` : Drug derived from living organism (proteins, antibodies, nucleic acids)
- `False` : Not a biologic

- Biologics face **biosimilar** competition with typically slower/different erosion patterns

`small_molecule`:

- `True`: Low molecular weight, chemically synthesized compound
- `False`: Not a small molecule (often when `biological` is `True`)
- Small molecules typically face many inexpensive generics with faster, deeper erosion

Example Data

country	brand_name	ther_area	hospital_rate	main_package	biological	small_molecule
COUNTRY_0024	BRAND_1143	Sensory_organs	0.09	EYE DROP	False	True
COUNTRY_0024	BRAND_1865	Muscoskeletal_Rheu...	92.36	INJECTION	False	False
COUNTRY_0024	BRAND_2F6C	Antineoplastic_and...	0.01	INJECTION	True	False

Note: Missing values (`nan`) exist in some columns and must be handled during preprocessing.

7.6 Erosion Buckets Distribution

The 340 test observations are distributed across **both scenarios** and **both erosion buckets**.

Dataset	Observations	Contains
Training	1,953	All erosion profiles (B1 and B2)
Test - Scenario 1	228	Mix of B1 (high) + B2 (low/medium) erosion
Test - Scenario 2	112	Mix of B1 (high) + B2 (low/medium) erosion

Important:

- Both scenarios contain **both B1 and B2 cases**
- The proportion/structure is consistent between both scenarios
- Models must handle **both kinds of erosion dynamics** within each scenario

7.7 Additional Data Guidelines

Guideline	Details
Granularity	Monthly level, starting from brand launch or first available data
Data Usage	You are free to design models that can be applied to both scenarios. Note that true post-entry volumes of the test set are never visible and cannot be used for training.
Bucket Labels	Not provided; derive using Mean Generic Erosion formula
Modeling Freedom	Any approach/model allowed; **explainability and simplicity valued**
Volume Units	"Number of units sold". Units may differ by product; treat series as comparable only after normalization per series.
Categorical Variables	Assumed constant over time
Missing Values	Present in some columns; preprocessing strategy is participant's choice (keep, impute, or drop)

7.8 Helper / Example Files

The organizers provide several helper files to assist with validation and submission:

File	Purpose
`auxiliar_metric_computation_example.csv`	Small toy example showing the structure of the auxiliary file
`submission_template.csv`	Empty template for all test `(country, brand_name, months_postgx)` combinations
`submission_example.csv`	Same structure as template, populated with dummy volumes (e.g., all zeros) to illustrate exact CSV format
`metric_calculation.py`	Official Python implementation of Metric 1 (Phase 1-a) and Metric 2 (Phase 1-b) for local validation

7.8.1 Auxiliary File Structure (`auxiliar_metric_computation.csv`)

You must compute an auxiliary file containing:

Column	Description
`country`	Market identifier
`brand_name`	Brand identifier
`avg_vol`	Average monthly volume (12 months before generic entry)
`bucket`	1 (high erosion) or 2 (medium/low erosion)

> **Scope**: This auxiliary file is used **only for local validation on the training data** together with `metric_calculation.py` . It is **not submitted** to the competition platform; the organizers will use their own internal auxiliary file for the hidden test set.

7.8.2 Metric Calculation Script (`metric_calculation.py`)

The official Python script provides two functions for local validation:

```
compute_metric1(df_actual, df_pred, df_aux)  # Phase 1-a (Scenario 1)
compute_metric2(df_actual, df_pred, df_aux)  # Phase 1-b (Scenario 2)
```

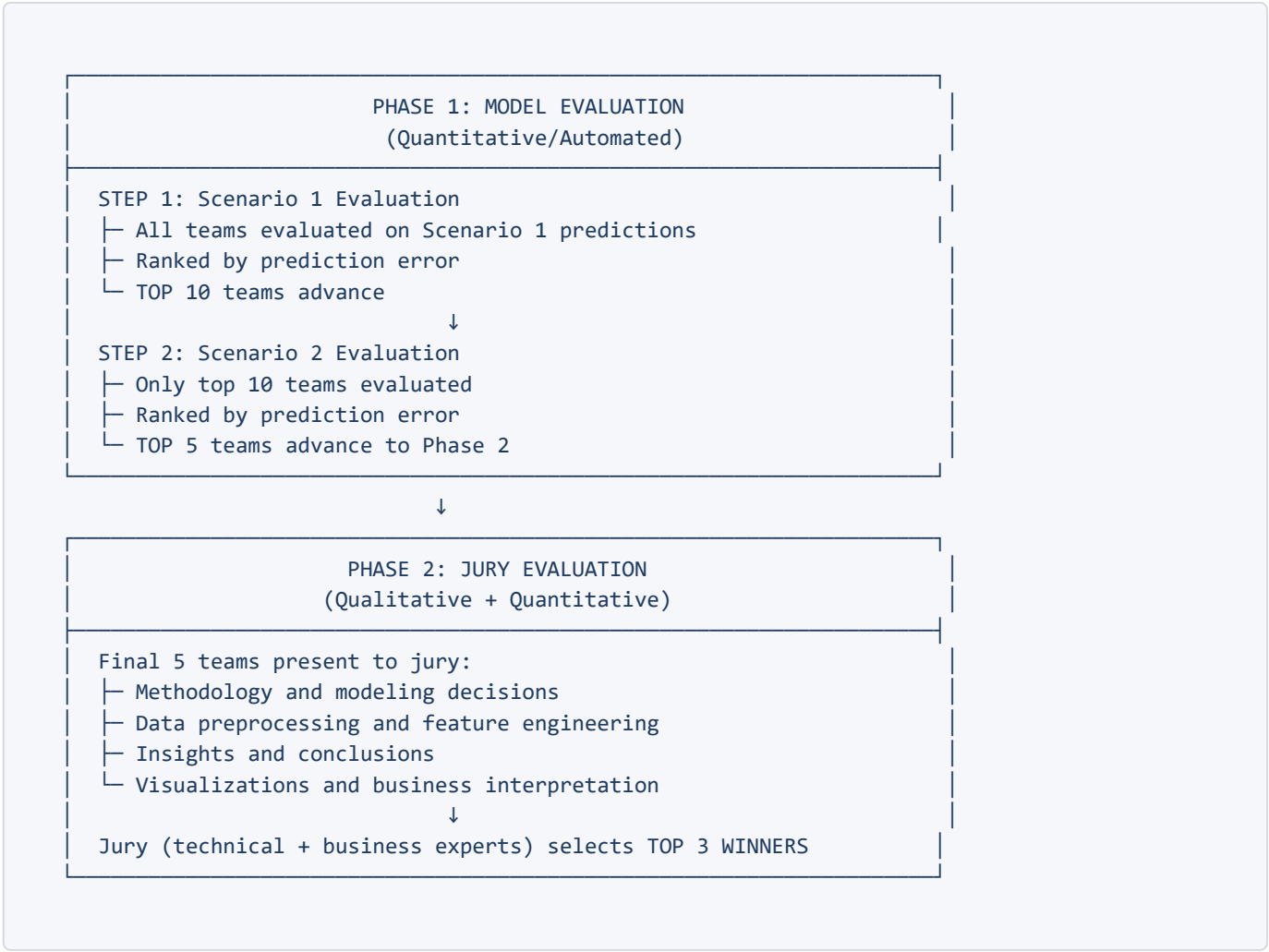
Important Implementation Details:

1. **Column Requirements**: For local validation, both `df_actual` and `df_pred` must contain the columns `country` , `brand_name` , `months_postgx` , and `volume` .
2. **Scenario Detection**: The script assumes that for Scenario 1, the slice you pass for each series starts at month 0 (`start_month == 0`). For Scenario 2, it assumes the slice starts at month 6 (`start_month == 6`). Any other rows in `df_actual` are ignored.

---

8. Evaluation Framework

8.1 Two-Phase Evaluation Process



8.2 Evaluation Summary

Phase	Teams Evaluated	Criterion	Teams Advancing
Phase 1-a	All	Scenario 1 accuracy	Top 10
Phase 1-b	Top 10	Scenario 2 accuracy	Top 5
Phase 2	Top 5	Jury presentation	Top 3 (Winners)

8.3 Phase 2: Jury Evaluation

The **five finalist teams** present their work to a jury composed of technical and business experts.

Presentation Requirements

Component	Description
Methodology	Data preprocessing, feature engineering, modeling choices, validation strategy
Modeling Decisions	Justification for choices made
Insights	Patterns discovered, drivers of erosion identified, especially for high-erosion markets
Conclusions	Business implications, recommendations, limitations

Jury Composition

- **Technical experts**: Data scientists, statisticians

- **Business experts**: Finance, commercial, market access professionals

Evaluation Criteria

- Technical robustness and soundness
- Interpretability and clarity
- Business relevance and practical applicability
- Quality of visualizations and communication

---

9. Evaluation Metrics: Detailed Mathematical Formulas

9.1 Metric Design Philosophy

The Prediction Error (PE) metric is designed to capture three key business dimensions:

Dimension	How Captured	Why Important
Erosion Severity	Bucket weighting (B1 × 2, B2 × 1)	High-erosion cases are most critical for business
Time Sensitivity	Period-specific weights (early months weighted highest)	Early erosion dynamics drive immediate business decisions
Pattern Accuracy	Monthly error terms	Captures month-to-month shape, not just cumulative totals

Normalization: All error components are normalized by Avg_j (pre-generic average monthly volume) for cross-series comparability. This ensures fair comparison across large and small brands.

> **Note**: If a series has $\text{Avg}_j = 0$ or is missing (NaN), its PE is set to NaN and is excluded from the metric computation.

9.2 Scenario 1 Prediction Error Formula

Participants predict months 0–23 **without any post-generic actuals**.

Four Error Components

Component	Weight	Period	Description
Monthly error	20%	Months 0–23	Month-by-month accuracy across full horizon
Accumulated error	**50%**	Months 0–5	Total volume accuracy in critical first 6 months
Accumulated error	20%	Months 6–11	Total volume accuracy in months 6–11
Accumulated error	10%	Months 12–23	Total volume accuracy in second year

Total: 20% + 50% + 20% + 10% = 100%

Complete Formula (4.1)

$$PE_j = 0.2 \cdot T_1 + 0.5 \cdot T_2 + 0.2 \cdot T_3 + 0.1 \cdot T_4$$

Term T_1 — Monthly Error (20%): $$T_1 = \frac{\sum_{i=0}^{23} \left| Y^{act}*{j,i} - Y^{pred}*{j,i} \right|}{24 \cdot Avg_j}$$

Term T_2 — Accumulated Error Months 0–5 (50%): $$T_2 = \frac{\left| \sum_{i=0}^{5} Y^{act}*{j,i} - \sum*{i=0}^{5} Y^{pred}_{j,i} \right|}{6 \cdot Avg_j}$$

Term T_3 — Accumulated Error Months 6–11 (20%): $$T_3 = \frac{\left| \sum_{i=6}^{11} Y^{act}*{j,i} - \sum*{i=6}^{11} Y^{pred}_{j,i} \right|}{6 \cdot Avg_j}$$

Term T_4 — Accumulated Error Months 12–23 (10%): $$T_4 = \frac{\left| \sum_{i=12}^{23} Y^{act}*{j,i} - \sum*{i=12}^{23} Y^{pred}_{j,i} \right|}{12 \cdot Avg_j}$$

Where:

- $Y^{act}_{j,i}$ = actual volume for series j in month i
- $Y^{pred}_{j,i}$ = predicted volume for series j in month i
- Avg_j = pre-generic average monthly volume (average of 12 months before entry)

9.3 Scenario 2 Prediction Error Formula

Participants have 6 actual post-entry months (0–5) available; predictions required for months 6–23 only.

Three Error Components

Component	Weight	Period	Description
Monthly error	20%	Months 6–23	Month-by-month accuracy across 18-month horizon
Accumulated error	**50%**	Months 6–11	Total volume accuracy in first 6 forecast months
Accumulated error	30%	Months 12–23	Total volume accuracy in second year

Total: 20% + 50% + 30% = 100%

Complete Formula (4.2)

$$PE_j = 0.2 \cdot T_1 + 0.5 \cdot T_2 + 0.3 \cdot T_3$$

Term T_1 — Monthly Error (20%): $$T_1 = \frac{\sum_{i=6}^{23} \left| Y^{act}*{j,i} - Y^{pred}*{j,i} \right|}{18 \cdot Avg_j}$$

Term T_2 — Accumulated Error Months 6–11 (50%): $$T_2 = \frac{\left| \sum_{i=6}^{11} Y^{act}*{j,i} - \sum*{i=6}^{11} Y^{pred}_{j,i} \right|}{6 \cdot Avg_j}$$

Term T_3 — Accumulated Error Months 12–23 (30%): $$T_3 = \frac{\left| \sum_{i=12}^{23} Y^{act}*{j,i} - \sum*{i=12}^{23} Y^{pred}_{j,i} \right|}{12 \cdot Avg_j}$$

9.4 Final Score Aggregation

After computing PE_j for each country-brand, the final competition score aggregates across buckets.

Formula (4.3): Bucket-Weighted Final Score

$$PE = \frac{2}{n_{B1}} \sum_{j=1}^{n_{B1}} PE_{j,B1} + \frac{1}{n_{B2}} \sum_{j=1}^{n_{B2}} PE_{j,B2}$$

Where:

- n_{B1} = number of test observations in Bucket 1 (high erosion)
- n_{B2} = number of test observations in Bucket 2 (low/medium erosion)
- $PE_{j,B1}$ = prediction error for brand j in Bucket 1
- $PE_{j,B2}$ = prediction error for brand j in Bucket 2

Interpretation

Term	Meaning

Term	Meaning
$\frac{1}{n_{B1}} \sum PE_{j,B1}$	Average error across high-erosion brands
$\frac{1}{n_{B2}} \sum PE_{j,B2}$	Average error across low-erosion brands
Factor **2** on B1	**Bucket 1 counts twice as much as Bucket 2**

Rationale: High-erosion cases (Bucket 1) are more business-critical, so they receive double weight in the final score.

9.5 Metric Interpretation Guide

Per-Brand Error (PE_j) Interpretation

PE_j Value	Interpretation
0	Perfect prediction
0.1 - 0.3	Excellent predictions
Close to 0.5	Moderate accuracy
Close to 1	Average error ≈ baseline monthly volume
> 1	Poor predictions; errors exceed typical pre-LOE volumes

Final Score (PE) Ranges

Final PE	Interpretation
0	All predictions perfect
3	All $PE_j = 1$ (since 2×1 + 1×1 = 3)
> 3	Some individual errors exceed 1

Goal: Minimize final PE score. **Lower is better.**

---

10. Submission Platform and Process

A dedicated **submission platform** is used for uploading predictions, computing metrics, and generating the leaderboard.

10.1 Platform Access

- Log in using your **team's username and password**
- Access link provided in submission instructions document

First action after login:

1. Navigate to options on the right
2. Click on **"Profile"**
3. Select **"Change password"**
4. Update your credentials

10.2 Uploading Submissions

1. Go to the left-hand menu
2. Click on **"Dashboard/Panel"**
3. Click on **"Checkpoint"**
4. Use the **upload button** to submit your file

If you see an error message:

- The file structure is incorrect
- Check columns, formatting, and missing fields
- Adjust and resubmit

Once a valid file is uploaded:

- Your team appears in the ranking
- Each team shows only its **best solution**
- Ranking updates when a new valid submission improves the score

10.3 Submission File Format

The submission file (`submission_template.csv`) must contain:

Column	Description
`country`	Market identifier
`brand_name`	Brand identifier
`months_postgx`	Forecast month (0–23 for Scenario 1, 6–23 for Scenario 2)
`volume`	**Your predicted volume**

Example:

```
country,brand_name,months_postgx,volume
COUNTRY_9891,BRAND_3C69,0,50000.0
COUNTRY_9891,BRAND_3C69,1,45000.0
COUNTRY_9891,BRAND_3C69,2,42000.0
```

10.4 Public vs Private Test Set

Test Set	Percentage	Usage
Public	30%	Used for online leaderboard during competition
Private	70%	Used for final evaluation after competition ends

Important: The public leaderboard score during the competition is based on only 30% of the test data. Final results use the complete test set.

10.5 Submission Limits

Limit	Details
Maximum submissions	3 per team every 8 hours
Recommendation	Make a test submission within the first few hours to verify format

---

11. Communication and Collaboration

All **communication** between teams and mentors takes place through **Microsoft Teams**.

11.1 Channel Structure

Mentoring Channel (Private)

- **Access:** Only members of your specific team + assigned mentors
- **Purpose:**
  - Ask questions
  - Request feedback
  - Schedule and conduct mentoring meetings

Novartis Datathon Channel (General)

Contains:

- **General sub-channel:** Announcements and organizational updates (only mentors can post)
- **Files tab:** Resources and submission documentation

11.2 Mentoring Sessions

- Mentors will initiate meetings in the private channel
- Join meetings by clicking the **"Meet"** button at the agreed time
- Use this for feedback on approach, clarifications, and guidance

11.3 File Resources

In the **Files tab** of the Datathon channel:

Resource	Description
"Data for Participants" folder	Dataset files
"Submissions" folder	Metrics documentation, submission instructions, data format examples
Slide templates	For finalist presentations

---

12. Timeline and Final Event

12.1 Final Selection Process

Time	Action
Sunday 9:30 AM	"Select your final option" feature activated
Sunday 10:30 AM	Datathon ends; no more submissions
After close	Final evaluation on complete test set
Following	Top 10 by Scenario 1 score published
Following	Top 5 by Scenario 2 score published (finalists)

12.2 Finalist Requirements

Only the **5 finalist teams** must:

Presentation Upload

- **Deadline:** 12:00 PM (noon)
- **Location:** Private channel in Microsoft Teams
- **Content:** Summary of methodology, models, key insights, results, conclusions
- **Format:** Use provided slide template

- **Naming:** Follow specified naming convention

Code Submission

- Upload the code used to generate your final submission results
- Same deadline and location as presentation

Final Event Schedule

Time	Activity
1:00 PM	Finalist presentations to jury and audience
2:30 PM	Jury deliberation and winner announcement

---

13. Modeling Recommendations and Best Practices

13.1 Data Preprocessing

Task	Recommendations
Missing Values	Choose strategy: imputation (mean/median, forward-fill), model-based handling, or dropping
Normalization	Normalize volumes per series using pre-generic baseline
Bucket Derivation	Calculate MGE for training data to identify Bucket 1 vs Bucket 2 cases
Volume Units	Account for different units across country-brand pairs; consider per-series normalization
Outlier Treatment	Document and justify any outlier handling decisions

13.2 Feature Engineering

Feature Type	Examples
Time-based	`months_postgx`, seasonality indicators, time since launch
Competition	`n_gxs`, binary generic presence, months since first generic, change in `n_gxs`
Product characteristics	One-hot encoded `ther_area`, `main_package`, `biological`, `small_molecule`
Distribution	`hospital_rate`, potential interactions with competition
Historical patterns	Pre-generic trend, volatility, growth rate
Derived metrics	Pre-generic average, normalized volumes, erosion bucket

13.3 Model Design Priorities

Priority	Rationale	Metric Impact
Early erosion accuracy	First 6 months most critical for business decisions	50% weight on first evaluated period
High-erosion case performance	Bucket 1 is the primary business focus	Bucket 1 has 2× importance
Monthly pattern matching	Captures actual shape of erosion curve	20% weight on per-month errors
Cumulative accuracy	Ensures total volume predictions are reasonable	30-50% weight on period totals
Interpretability	Required for jury presentation	Valued in Phase 2 evaluation

Priority	Rationale	Metric Impact
Simplicity	Easier to explain and justify	Explicitly valued by organizers

13.4 Scenario-Specific Strategies

Scenario 1 (No post-entry data)

- Rely on pre-generic patterns and product characteristics
- Learn typical erosion shapes from training data
- Consider classification approach (bucket prediction) alongside regression
- Feature importance from similar historical cases
- May benefit from panel/hierarchical models leveraging across-series patterns

Scenario 2 (6 months post-entry data)

- Update forecasts using observed early erosion behavior
- Calibrate predictions based on actual erosion trajectory
- Leverage early months to identify erosion bucket
- May benefit from conditional modeling given early behavior
- Consider blending pre-generic features with observed post-generic patterns

---

14. Appendix: Mathematical Formulas and Quick Reference

Core Erosion Metrics

Formula	Equation
Mean Generic Erosion	$\text{MGE} = \frac{1}{24} \sum_{i=0}^{23} \text{vol}_i^{\text{norm}}$
Normalized Volume	$\text{vol}_i^{\text{norm}} = \frac{\text{Vol}_i}{\text{Avg}_j}$
Baseline Volume	$\text{Avg}*j* = *frac{1}{12}* *sum*{i=-12}^{-1} Y_{j,i}^{\text{act}}$

Prediction Error Metrics

Formula	Application
PE Scenario 1 (4.1)	Evaluates Months 0-23 predictions
PE Scenario 2 (4.2)	Evaluates Months 6-23 predictions
Final PE (4.3)	Bucket-weighted aggregation

Bucket Thresholds

Bucket	MGE Condition	Description
Bucket 1 (B1)	$0 \leq \text{MGE} \leq 0.25$	High erosion (≥75% loss)
Bucket 2 (B2)	$0.25 < \text{MGE} \leq 1$	Medium/low erosion

Weight Summary

Component	Scenario 1	Scenario 2
Monthly errors	20%	20%

Component	Scenario 1	Scenario 2
Early cumulative	50% (M0-5)	50% (M6-11)
Mid cumulative	20% (M6-11)	—
Late cumulative	10% (M12-23)	30% (M12-23)

Bucket Weighting

Bucket	Weight
Bucket 1	**2×**
Bucket 2	1×

Quick Reference Summary

Item	Value
Total country-brand combinations	2,293
Training observations	1,953
Test observations (Scenario 1)	228
Test observations (Scenario 2)	112
Forecast horizon (Scenario 1)	24 months (M0-M23)
Forecast horizon (Scenario 2)	18 months (M6-M23)
Top teams advancing from Phase 1-a	10
Finalist teams	5
Winning teams	3
Bucket 1 threshold	MGE ≤ 0.25
Bucket 1 weight multiplier	2×
Highest component weight	50% (early cumulative)
Public test set percentage	30%
Private test set percentage	70%

Document Version: Final Consolidated | Novartis Datathon 2025 | Barcelona Digital Finance Hub