Novartis Datathon 2025: Complete Comprehensive Guide

Table of Contents

---

1. Introduction and Context

1.1 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

1.2 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

1.3 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

---

2. The Barcelona Digital Finance Hub

2.1 History and Growth

Year	Team Size
2018	10
2019	19
2020	28
2021	34
2022	40
2023	50
2024	53
2025	55

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

2.2 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

2.3 Team Diversity

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

2.4 Educational Background

Discipline	Percentage
Mathematics & Statistics	24%
Computer Science	20%

Discipline	Percentage
Economics	20%
Physics & Others	19%
Engineering	17%

Additional qualifications:

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

2.5 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

2.6 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

---

3. Business Background: The Generics Problem

3.1 Patents and Loss of Exclusivity (LOE)

When a pharmaceutical company develops a new drug, it receives a **patent** providing exclusive rights for a limited period (typically 20 years from filing). Key concepts:

- **Patent Protection**: During exclusivity, no other company can legally manufacture or sell the same drug
- **Loss of Exclusivity (LOE)**: When patent protection expires, allowing generic manufacturers to enter the market
- **Generic Entry**: The moment when generic versions of the drug become available

3.2 What is a Generic Drug?

A generic drug is **therapeutically equivalent** to the brand-name medication in terms of:

- Dosage form (tablet, capsule, injectable)
- Strength (amount of active ingredient)

- Route of administration (oral, intravenous)
- Quality and performance
- Intended use

Key Differences:

- Inactive ingredients (fillers, binders, colorants) may differ
- These differences do not affect therapeutic effect

3.3 Bioequivalence

Generic manufacturers must demonstrate **bioequivalence** to the original product by comparing pharmacokinetic properties:

- **Absorption**: Rate and extent of drug entry into bloodstream
- **Distribution**: Spread through body tissues
- **Metabolism**: How the body transforms the drug
- **Elimination**: Excretion via kidneys or liver

This reduced regulatory requirement allows generics to be developed at **significantly lower costs**.

3.4 Consequences of Generic Entry

1. **Increased Competition**: Multiple manufacturers drive prices downward
2. **Improved Affordability**: Lower prices for patients and healthcare systems
3. **Greater Access**: More patients can access treatment
4. **Prescribing Changes**: Healthcare professionals may prescribe generics; pharmacies may substitute

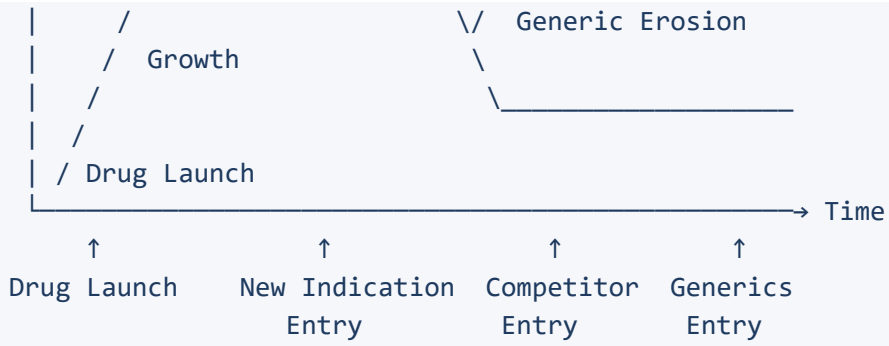
3.5 Example: Diovan (Valsartan)

- **Brand**: Diovan (Novartis)
- **Active Ingredient**: Valsartan (angiotensin II receptor blocker)
- **Indications**: Hypertension, heart failure
- **Patent Expiry**: 2012
- **Result**: Multiple generic manufacturers entered, significantly increasing competition and reducing prices

---

4. Drug Lifecycle and Generic Erosion

4.1 Drug Sales Lifecycle Phases



Phase Descriptions:

1. **Launch Phase**
    - Drug first introduced to market
    - Sales volume initially low
    - Awareness and adoption developing
2. **Growth Phase**
    - Product gains market acceptance
    - Sales volumes increase rapidly
    - More doctors prescribe, more patients adopt
3. **Maturity Phase**
    - Sales stabilize at high level
    - Market becomes saturated
    - Growth slows as most potential prescribers already use the drug
4. **Competitive Pressure**
    - New competitor drugs enter
    - Growth slows or begins to decline
    - Market share pressure increases
5. **Loss of Exclusivity (LOE) and Generic Entry**
    - Patent expires
    - Generic products enter the market
    - **Sharp drop in sales volume** = **Generic Erosion**

4.2 The Datathon Focus

The Datathon specifically focuses on the **post-LOE period**—the critical phase when generics enter and sales decline. This is where forecasting is most valuable for business decisions.

---

5. Mean Generic Erosion: Mathematical Definition

5.1 Core Concept

Mean Generic Erosion quantifies how much a branded drug's sales fall after generic competitors enter. It is defined as the **mean of normalized volumes** during the **24 months following generic entry**.

5.2 Mathematical Formulas

Formula 1: Mean Generic Erosion

$$\text{Mean Generic Erosion} = \frac{1}{24} \sum_{i=0}^{23} \text{vol}_i^{\text{norm}}$$

Where:

- i = month index after generic entry (0 to 23)
- $\text{vol}_i^{\text{norm}}$ = normalized volume in month i

Formula 2: Normalized Volume

$$\text{vol}_i^{\text{norm}} = \frac{\text{Vol}_i}{\text{Avg}_j}$$

Where:

- Vol_i = actual sales volume in month i after generic entry
- Avg_j = baseline volume for drug j

Formula 3: Pre-Generic Average (Baseline)

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

Where:

- $Y_{j,i}^{\text{act}}$ = actual sales volume for drug j in month i
- Sum runs over the **12 months before generic entry** ($i = -12$ to -1)

5.3 Interpretation

Mean Generic Erosion	Interpretation
Close to 1	Little erosion; post-generic volume ≈ pre-generic average
0.5	Moderate erosion; 50% of pre-generic volume retained
Close to 0	Severe erosion; sales collapsed
> 1	Unusual; volume increased (very rare)

5.4 Step-by-Step Calculation

1. **Compute baseline** (Avg_j): Average monthly volume over 12 months before generic entry
2. **Normalize each post-entry month**: Divide each month's volume by the baseline
3. **Calculate mean**: Average the 24 normalized post-entry volumes

This normalization allows **comparison across drugs and markets** regardless of absolute sales magnitude.

---

6. Erosion Classification and Buckets

6.1 Three Conceptual Erosion Patterns

1. **Low Erosion**
    - Minimal impact after generic entry
    - Volume remains relatively stable
    - Mean normalized erosion **close to 1**
2. **Medium Erosion**
    - Moderate decline in sales
    - Mean erosion **between 0 and 1**
    - Partial loss of volume
3. **High Erosion**
    - Sharp drop in volume
    - Mean normalized erosion **close to 0**
    - Severe loss of market share

6.2 Datathon Bucket Classification

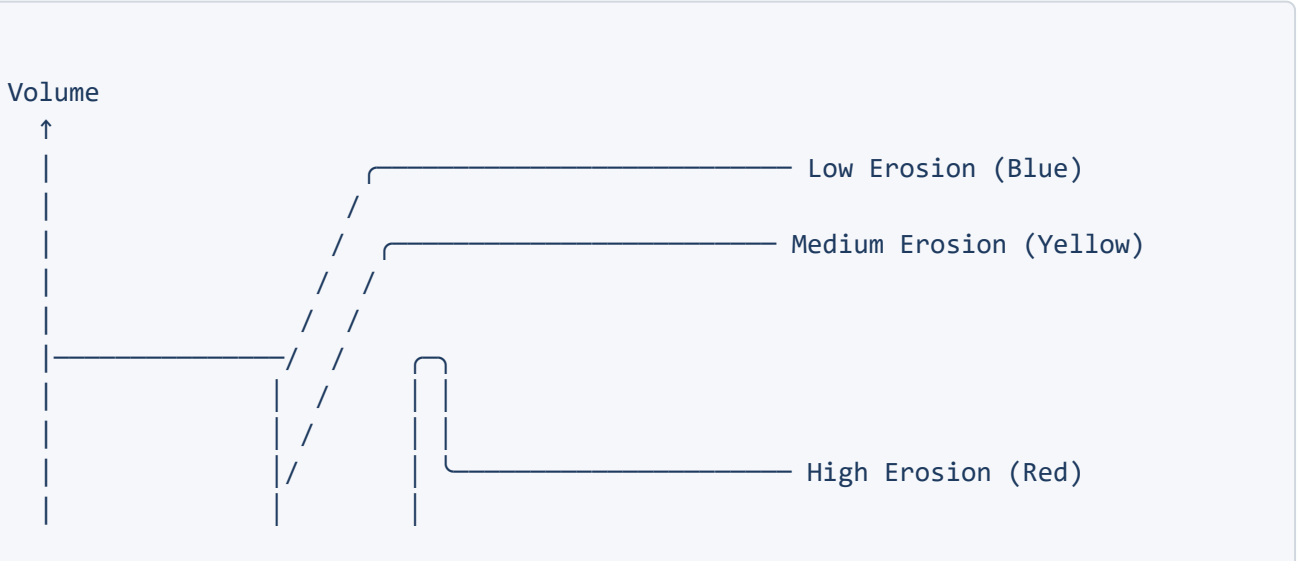
For the Datathon, drugs are classified into **two numerical buckets**:

Bucket	Mean Erosion Range	Description	Datathon Focus
Bucket 1 (B1)	[0, 0.25]	High erosion drugs	**PRIMARY FOCUS**
Bucket 2 (B2)	(0.25, 1]	Medium/Low erosion	Secondary

Note: The interval notation:

- `[0, 0.25]` = 0 to 0.25, inclusive
- `(0.25, 1]` = greater than 0.25, up to and including 1

6.3 Visual Representation of Erosion Patterns



```
└──────────────┼──────────┼──────────────────────→ Month_num
           Generics
           entry date
               │
  ← Pre-Gx     │    Post-Gx →
```

6.4 Business Importance

Bucket 1 (High Erosion) receives special attention because:

- More financially critical and risky
- Harder to predict accurately
- Most important for strategic planning
- **Double-weighted** in evaluation metrics

---

7. The Datathon Challenge

7.1 Objective

Forecast volume erosion after generic entry over a **24-month horizon** from the generic entry date.

7.2 Two Forecasting Scenarios

Scenario 1: Right After Generic Entry

Aspect	Details
Time Position	Immediately at generic entry (month 0)
Available Data	Pre-generic history only
Post-Entry Actuals	None
Forecast Required	Month 0 to Month 23 (24 months)
Business Simulation	Early planning at LOE with no market data

Scenario 2: Six Months After Generic Entry

Aspect	Details
Time Position	6 months after generic entry
Available Data	Pre-generic history + 6 months post-entry
Post-Entry Actuals	Months 0–5 provided
Forecast Required	Month 6 to Month 23 (18 months)
Business Simulation	Mid-course forecast update with early erosion data

7.3 Technical Challenge

Build forecasting models that work in **two different information settings**:

- **Scenario 1**: Pure ex-ante prediction
- **Scenario 2**: Updated forecast with partial post-entry data

7.4 Business Challenge

Teams must also demonstrate:

- **Deep exploratory analysis** of data preprocessing
- **Special focus on high-erosion cases** (Bucket 1)
- **Clear visualizations** for business stakeholders
- **Justification** of modeling approaches
- **Business-oriented** interpretation of results

---

8. Data Structure and Datasets

8.1 Overview

Dataset	Observations	Description
Training Set	1,953	Country-brand combinations with full history
Test Set	340	For evaluation (228 Scenario 1, 112 Scenario 2)
Total	2,293	Country-brand combinations that experienced generic entry

8.2 Training Set Details

Each training observation includes:

- Up to **24 months of volume data before** generic entry
- Up to **24 months of volume data after** generic entry

This allows analysis of **pre- and post-entry dynamics**.

8.3 Test Set Breakdown

Scenario	Observations	Percentage	Forecast Range
Scenario 1	228	~67%	Month 0 to 23
Scenario 2	112	~33%	Month 6 to 23

Both scenarios contain observations from **both Bucket 1 and Bucket 2**.

8.4 Three Datasets Provided

> **File Naming Convention**: Throughout this guide, when we refer to a dataset conceptually (e.g., "the volume dataset"), we mean:

> - `df_volume_train.csv` for training data
> - `df_volume_test.csv` for test data
>
> The same pattern applies to generics (`df_generics_train.csv` / `df_generics_test.csv`) and medicine info (`df_medicine_info_train.csv` / `df_medicine_info_test.csv`).

8.4.1 Volume Dataset (`df_volume_train.csv` / `df_volume_test.csv`)

Purpose: Core time series of monthly sales volumes

Column	Description	Example
`country`	Market identifier	`COUNTRY_B6AE`
`brand_name`	Brand identifier	`BRAND_1C1E`
`month`	Calendar month	`Jul` , `Aug` , etc.
`months_postgx`	Months relative to generic entry	`-24` to `23`
`volume`	**TARGET VARIABLE** - Units sold	`272594.39`

Key Notes on `months_postgx` :

- `0` = Month of generic entry
- Negative values = Months **before** entry
- Positive values = Months **after** entry

Sample 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
```

8.4.2 Generics Dataset (`df_generics_train.csv` / `df_generics_test.csv`)

Purpose: Information about generic competition over time

Column	Description	Example
`country`	Market identifier	`COUNTRY_B6AE`
`brand_name`	Brand identifier	`BRAND_DF2E`
`months_postgx`	Months after generic entry	`0` , `1` , `2` , etc.
`n_gxs`	Number of generics on market	`0.0` , `1.0` , `2.0`

Key Notes:

- Number of generics is **time-varying**
- New generics can enter; some may exit
- Crucial feature for modeling competitive pressure

Sample 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
```

8.4.3 Medicine Information Dataset (`df_medicine_info_train.csv` / `df_medicine_info_test.csv`)

Purpose: Static characteristics of each drug

Column	Type	Description
`country`	Categorical	Market identifier
`brand_name`	Categorical	Brand identifier
`ther_area`	Categorical	Therapeutic area (e.g., `Sensory_organs` , `Nervous_system`)
`hospital_rate`	Numeric	Percentage delivered via hospitals (0-100)
`main_package`	Categorical	Dosage form (`PILL` , `INJECTION` , `EYE DROP`)
`biological`	Boolean	True if biologic product (proteins, antibodies)
`small_molecule`	Boolean	True if low molecular weight compound

Sample 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...    92.36          INJECTION
False       False
COUNTRY_0024 BRAND_2F6C  Antineoplastic...   0.01           INJECTION     True
False
```

Key Notes:

- Features are **time-invariant** (constant over time)
- `biological` and `small_molecule` are typically mutually exclusive

- Missing values exist (e.g., `hospital_rate` may be `NaN`)

8.5 Important Data Notes

1. **Volume Units**: May differ by country-brand (milligrams, packs, pills)
2. **Missing Values**: Some columns contain NaN; preprocessing required
3. **Bucket Information**: Not provided directly; must be computed using Mean Generic Erosion formula
4. **Data Usage**: All **training data** can be used to train models for both scenarios. Whether you use **test features** (without labels) for any unsupervised or semi-supervised step should comply with the competition's official rules; if in doubt, treat the test set purely as a hold-out evaluation set.
5. **Categorical Variables**: Can be assumed constant over time

8.6 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

8.6.1 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.

Note: The `auxiliar_metric_computation_example.csv` file demonstrates this structure with sample data.

8.6.2 Submission Files

- **`submission_template.csv`** : Contains all required `(country, brand_name, months_postgx)` combinations with an empty `volume` column for you to fill with predictions.
- **`submission_example.csv`** : Shows the exact required CSV format with dummy volume values.

8.6.3 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)
```

Expected Input DataFrames:

DataFrame	Required Columns
`df_actual`	`country` , `brand_name` , `months_postgx` , `volume`
`df_pred`	`country` , `brand_name` , `months_postgx` , `volume`
`df_aux`	`country` , `brand_name` , `avg_vol` , `bucket`

Internal Processing:

1. **Merges** `df_actual` and `df_pred` on `(country, brand_name, months_postgx)` , then merges with `df_aux`
2. **Computes** **`start_month`** per `(country, brand_name)` as the minimum `months_postgx` in that group
3. **Filters by scenario:**
   - Metric 1: Keeps only series where `start_month == 0` (predictions start at month 0)
   - Metric 2: Keeps only series where `start_month == 6` (predictions start at month 6)
4. **Groups** by `(country, brand_name, bucket)` and applies the per-series PE formula
5. **Aggregates** by bucket with the 2× / 1× weighting

> **Important**: Your validation and submission files must be aligned with the scenario they correspond to—predictions starting at month 0 for Scenario 1, and at month 6 for Scenario 2. The script uses the `start_month` filter to enforce this.

---

9. Evaluation Framework

9.1 Two-Phase Evaluation Process

```
┌──────────────────────────────────────────────────────────────────────┐
│                      PHASE 1: MODEL EVALUATION                       │
│                       (Quantitative/Automated)                       │
├──────────────────────────────────────────────────────────────────────┤
│  STEP 1: Scenario 1 Evaluation                                       │
```

```
│  ├─ All teams evaluated on Scenario 1 predictions                    │
│  ├─ Ranked by prediction error                                       │
│  └─ TOP 10 teams advance                                             │
│                              ↓                                       │
│  STEP 2: Scenario 2 Evaluation                                       │
│  ├─ Only top 10 teams evaluated                                      │
│  ├─ Ranked by prediction error                                       │
│  └─ TOP 5 teams advance to Phase 2                                   │
└──────────────────────────────────────────────────────────────────────┘
                              ↓
┌──────────────────────────────────────────────────────────────────────┐
│                    PHASE 2: JURY EVALUATION                          │
│                  (Qualitative + Quantitative)                        │
├──────────────────────────────────────────────────────────────────────┤
│  Final 5 teams present to jury:                                      │
│  ├─ Methodology and modeling decisions                               │
│  ├─ Data preprocessing and feature engineering                       │
│  ├─ Insights and conclusions                                         │
│  └─ Visualizations and business interpretation                       │
│                              ↓                                       │
│  Jury (technical + business experts) selects TOP 3 WINNERS           │
└──────────────────────────────────────────────────────────────────────┘
```

9.2 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)

---

10. Scoring Metrics: Detailed Mathematical Formulas

10.1 Common Elements

All metrics share these characteristics:

- **Normalized** by pre-generic average volume (Avg_j)
- **Bucket-weighted**: Bucket 1 weighted **twice** as much as Bucket 2
- Based on **absolute prediction errors**

10.2 Phase 1-a Metric (Scenario 1)

Per-Country-Brand Prediction Error (PE_j)

$$PE_j = 0.2\left(\frac{\sum_{i=0}^{23} |Y^{act}*{j,i} - Y^{pred}*{j,i}|}{24 \cdot Avg_j}\right) + 0.5\left(\frac{|\sum_{i=0}^{5} Y^{act}*{j,i} - \sum*{i=0}^{5} Y^{pred}_{j,i}|}{6 \cdot Avg_j}\right)$$

$$+ 0.2\left(\frac{|\sum_{i=6}^{11} Y^{act}*{j,i} - \sum*{i=6}^{11} Y^{pred}*{j,i}|}{6 \cdot Avg_j}\right) + *0.1\left(\frac{|\sum*{i=12}^{23} Y^{act}*{j,i} - \sum*{i=12}^{23} Y^{pred}_{j,i}|}{12 \cdot Avg_j}\right)$$

Where:

- $Y^{act}_{j,i}$ = Actual volume for brand j at month i
- $Y^{pred}_{j,i}$ = Predicted volume for brand j at month i
- Avg_j = Average monthly volume in 12 months before generic entry

Error Components Breakdown

Component	Months	Weight	Purpose
Monthly Error	0–23	20%	Detailed month-by-month accuracy
Accumulated Error	0–5	**50%**	Early erosion (MOST CRITICAL)
Accumulated Error	6–11	20%	Medium-term erosion
Accumulated Error	12–23	10%	Long-term tail

10.3 Phase 1-b Metric (Scenario 2)

Per-Country-Brand Prediction Error (PE_j)

$$PE_j = 0.2\left(\frac{\sum_{i=6}^{23} |Y^{act}*{j,i} - *Y^{pred}*{j,i}|}{18 \cdot Avg_j}\right) + 0.5\left(\frac{|\sum_{i=6}^{11} Y^{act}*{j,i} - \sum*{i=6}^{11} Y^{pred}_{j,i}|}{6 \cdot Avg_j}\right)$$

$$+ 0.3\left(\frac{|\sum_{i=12}^{23} Y^{act}*{j,i} - \sum*{i=12}^{23} Y^{pred}_{j,i}|}{12 \cdot Avg_j}\right)$$

Error Components Breakdown

Component	Months	Weight	Purpose
Monthly Error	6–23	20%	Month-by-month accuracy (18 months)
Accumulated Error	6–11	**50%**	Early forecast horizon (MOST CRITICAL)
Accumulated Error	12–23	30%	Long-term cumulative

10.4 Final Aggregated Metric

For both phases, the final Prediction Error is:

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

Where:

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

Key Insight: Bucket 1 is weighted **twice** as much as Bucket 2, reflecting its business importance.

10.5 Metric Interpretation

PE Value	Interpretation
0	Perfect predictions
< 1	Good predictions
≈ 1	Error magnitude equals pre-generic average
> 1	Poor predictions
> 3	Very poor predictions

> **Lower PE is better.** The leaderboard is sorted in ascending order of the prediction error.

Score Range:

- Minimum (all $PE_j = 0$): Final PE = 0
- If all $PE_j = 1$: Final PE = 3
- Maximum: Can exceed 3 if individual errors are very large

10.6 Metric Rationale

The metric captures three key dimensions:

1. **Generic Erosion Severity**: Via bucket-level weighting (Bucket 1 gets 2× weight)
2. **Temporal Importance**: Via differential weights across time periods (early months emphasized)
3. **Seasonality Effects**: Via monthly error terms ensuring short-term fluctuations count

---

11. Submission Process and Platform

11.1 Accessing the Platform

1. Log in with team username and password
2. **First action**: Change your password
    - Navigate to options → Profile → Change password

11.2 Uploading Submissions

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

If error appears: File structure is incorrect (wrong columns, improper formatting, missing fields)

11.3 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
```

11.4 Leaderboard and Test Set Split

Test Set Portion	Usage
Public (30%)	Online leaderboard during competition
Private (70%)	Final evaluation after competition ends

11.5 Submission Limits

- **Maximum**: 3 submissions every 8 hours
- **Recommendation**: Make a test submission within the first few hours to verify format

11.6 Final Selection Timeline

Time	Event
Sunday 9:30 AM	"Select your final option" activated
Sunday 10:30 AM	**Datathon officially ends**
After 10:30 AM	Final evaluation on complete test set

---

12. Communication and Collaboration

12.1 Microsoft Teams Channels

Mentoring Channel (Private)

- Access: Team members + assigned mentors only
- Purpose: Questions, feedback, mentoring meetings
- Meetings initiated by clicking "Meet" button

Novartis Datathon Channel (General)

Contains:

- **General sub-channel**: Announcements from mentors
- **Files tab**:
  - "Data for Participants" folder
  - "Submissions" folder with:
    - Official metric documentation (for cross-validation)
    - Detailed submission instructions
    - Example data formats (including an example submission)
  - Slide templates for final presentations

> **Mentoring Meetings**: Mentoring sessions are held via the private Teams channel. Mentors will initiate meetings using the "Meet" button, and team members join from there.

---

13. Timeline and Final Event

13.1 Final Day Schedule

Time	Event
9:30 AM	Final submission selection opens
10:30 AM	Competition ends; no more submissions
12:00 PM	Deadline for finalist presentations and code upload
1:00 PM	Finalist team presentations
2:30 PM	Jury deliberation and winner announcement

13.2 Finalist Requirements

Top 5 teams must:

1. **Prepare presentation** covering:
   - Methodology
   - Models and algorithms
   - Data insights and exploratory analysis
   - Key results and conclusions
2. **Upload materials** by 12:00 PM:
   - Presentation file (following naming convention)
   - Code used to generate final submission

13.3 Results Publication Order

1. **Top 10** results published (ordered by Scenario 1 score)
2. **Top 5** results published (ordered by Scenario 2 score)

3. **Top 3** winners announced by jury

---

14. Best Practices and Recommendations

14.1 Modeling Approach

Aspect	Recommendation
Model Choice	Any approach allowed; explainability and simplicity valued
Validation	Use train/validation splits mimicking scenarios
Focus	Prioritize Bucket 1 (high erosion) performance
Features	Leverage all three datasets through joins

> **Tools & Languages**: Any programming language or library is allowed as long as the final submission is a CSV in the required format and the approach is explainable to the jury.

14.2 Data Preprocessing

1. **Handle missing values**: Choose imputation, flags, or exclusion
2. **Normalize appropriately**: Consider brand size differences
3. **Feature engineering**: Use `n_gxs`, time features, drug characteristics
4. **Join datasets** on `(country, brand_name)` or `(country, brand_name, months_postgx)`

14.3 Bucket Computation

To compute buckets for training data:

1. Calculate Avg_j (12-month pre-generic average)
2. Calculate Mean Generic Erosion using Formula 1
3. Assign Bucket 1 if erosion $\in$ [0, 0.25], else Bucket 2

14.4 Submission Tips

1. **Test early**: Submit within first few hours to verify format
2. **Monitor submissions**: Track your leaderboard position
3. **Save best models**: You may need to select your final submission
4. **Document everything**: Needed for finalist presentation

14.5 Presentation Tips

- Use **visualizations** liberally
- Explain **preprocessing decisions** clearly
- Focus on **high-erosion cases** (Bucket 1)
- Connect technical choices to **business implications**
- Make insights **accessible to non-technical stakeholders**

14.6 Key Success Factors

1. **Accurate early month predictions** (months 0–5 or 6–11 have highest weights)
2. **Strong Bucket 1 performance** (2× weight in final metric)
3. **Clear, interpretable methodology** for jury presentation
4. **Business-oriented insights** connecting to generic erosion dynamics

---

Appendix: Quick Reference

A.1 Key Formulas

Formula	Expression
Mean Generic Erosion	$\frac{1}{24} \sum_{i=0}^{23} \frac{Vol_i}{Avg_j}$
Pre-Generic Average	$Avg_j = \frac{1}{12} \sum_{i=-12}^{-1} Y^{act}_{j,i}$
Final Metric	$PE = \frac{2}{n_{B1}} \sum PE_{j,B1} + \frac{1}{n_{B2}} \sum PE_{j,B2}$

A.2 Key Numbers

Item	Value
Total observations	2,293
Training set	1,953
Test set	340
Scenario 1 test	228
Scenario 2 test	112
Pre-generic window	12 months
Post-generic horizon	24 months
Bucket 1 range	[0, 0.25]
Bucket 2 range	(0.25, 1]
Bucket 1 weight	2×
Bucket 2 weight	1×

A.3 Scenario Comparison

Aspect	Scenario 1	Scenario 2
Position	At generic entry	6 months after
Known actuals	Pre-generic only	Pre-generic + months 0–5
Forecast range	Month 0–23	Month 6–23
Most weighted period	Months 0–5 (50%)	Months 6–11 (50%)

Aspect	Scenario 1	Scenario 2
Test observations	228	112

Document compiled from official Novartis Datathon 2025 materials. For questions, refer to the Microsoft Teams channels or your assigned mentors.