Exploratory Data Analysis Report

Novartis Datathon 2025 - Generic Erosion Forecasting

Generated: November 28, 2025
Dataset: 93,744 observations × 11 features
Brands: 1,953 unique brands across 54 countries
Time Range: Months -24 to +23 (relative to generic entry)

Table of Contents

---

1. Executive Summary

Key Findings

Metric	Value	Implication
Bucket 1 (High Erosion)	6.7% of brands (130)	Minority class but **2× weight** in evaluation
Bucket 2 (Low Erosion)	93.3% of brands (1,823)	Majority class, normal weight
Imbalance Ratio	14:1	Severe class imbalance
Bucket 1 Total Erosion	88.27% volume loss	From 0.65 to 0.08 normalized
Bucket 2 Total Erosion	40.17% volume loss	From 0.86 to 0.52 normalized
Competition at Entry	2.04 generics avg	Doubles to ~4.24 by month 23

⚠ Critical Observations

1. **Class Imbalance is SEVERE**: Only 6.7% of brands are Bucket 1, but they carry 2× weight
2. **Early Months Matter Most**: Bucket 1 brands lose 65%+ volume in first 6 months

3. **Competition Builds Gradually**: Generic competitors double over 24 months
4. **Therapeutic Area Matters**: Anti-infectives erode fastest, Sensory organs slowest
5. **Hospital Rate Impact**: High hospital rate (75-100%) shows faster erosion

---

2. Dataset Overview

2.1 Data Dimensions

```
Shape: 93,744 rows × 11 columns
Total unique brands: 1,953
Countries: 54
Months range: -24 to +23 (pre-entry to post-entry)
```

2.2 Key Variables

Variable	Type	Description
`country`	Categorical	54 unique countries (anonymized)
`brand_name`	Categorical	1,953 unique brands (anonymized)
`months_postgx`	Numeric	Months relative to generic entry (-24 to +23)
`volume`	Numeric	Monthly sales volume
`n_gxs`	Numeric	Number of generic competitors (0-43)
`ther_area`	Categorical	14 therapeutic areas
`biological`	Binary	1 if biological drug, 0 if small molecule
`hospital_rate`	Numeric	% of sales through hospital channel
`avg_vol` / `avg_j`	Numeric	Pre-entry average volume (baseline)
`vol_norm`	Numeric	Normalized volume (volume / avg_j)
`bucket`	Binary	1 = high erosion, 2 = low erosion

2.3 Missing Values Analysis

- **`time_to_50pct`**: Many brands (especially Bucket 2) never reach 50% erosion
- This results in `NaN` values for time_to_50pct metric
- **Recommendation**: Use median imputation or create binary flag for "never reached 50%"

---

3. Bucket Distribution Analysis

3.1 Distribution Summary

Bucket	Count	Percentage	Weight	Effective Weight
Bucket 1 (High Erosion)	130	6.7%	2×	13.4%
Bucket 2 (Low Erosion)	1,823	93.3%	1×	86.6%

3.2 Bucket Definitions

- **Bucket 1**: Mean normalized volume ≤ 0.25 in months 18-23 (75%+ volume loss)
- **Bucket 2**: Mean normalized volume > 0.25 in months 18-23

3.3 Imbalance Analysis

```
Imbalance Ratio: 14.02:1 (Bucket 2 : Bucket 1)
```

⚠ **CRITICAL**: This severe imbalance requires special handling:

- Standard models will bias toward Bucket 2
- Bucket 1 predictions carry 2× weight in scoring
- Optimizing only for accuracy will hurt competition score

---

4. Erosion Curves by Bucket

4.1 Bucket 1 Erosion Profile (High Erosion)

Metric	Month 0	Month 6	Month 12	Month 23
Mean Vol Norm	0.648	0.162	0.108	0.076
Std Dev	0.285	0.107	0.126	0.070

Key Observations:

- **Initial Drop**: 35% erosion by month 0 (from avg baseline of 1.0)
- **First 6 Months**: Drops from 0.65 to 0.16 (75% of total erosion!)
- **Stabilization**: Levels off around 0.08 by month 12
- **Final Level**: ~7.6% of original volume remains

4.2 Bucket 2 Erosion Profile (Low Erosion)

Metric	Month 0	Month 6	Month 12	Month 23
Mean Vol Norm	0.864	0.697	0.582	0.517
Std Dev	0.171	0.249	0.255	0.287

Key Observations:

- **Initial Drop**: Only 14% erosion by month 0
- **Gradual Decline**: Steady erosion over 24 months
- **Higher Variability**: Std dev increases over time
- **Final Level**: ~52% of original volume remains

4.3 Monthly Erosion Rates

Bucket	Mean Monthly Decline	Total 24-Month Loss
Bucket 1	-2.49%	88.27%
Bucket 2	-1.51%	40.17%

Bucket 1 erodes 65% faster per month than Bucket 2

---

5. Competition Impact Analysis

5.1 Generic Competitor Trajectory

Time Point	Mean n_gxs	Std Dev
Month 0 (Entry)	2.04	3.45
Month 6	3.54	4.47
Month 12	4.24	4.76
Month 23	4.24	4.76

Pattern: Competition builds rapidly in first 12 months, then stabilizes.

5.2 Volume vs Number of Competitors

n_gxs	Mean Vol Norm	Observation
0	0.785	No competition yet
1	0.633	First generic entry
2	0.577	Two competitors
3	0.554	Three competitors
4	0.519	Four competitors
5	0.497	Five competitors
6	0.436	Six competitors
10+	~0.45	Diminishing returns

Key Insight: Volume drops rapidly with first 6 competitors, then plateaus.

5.3 Competition Impact Curve

```
n_gxs:    0 → 1 → 2 → 3 → 4 → 5 → 6 → 10+
vol_norm: 0.78 → 0.63 → 0.58 → 0.55 → 0.52 → 0.50 → 0.44 → ~0.45
          ↓ 19%  ↓ 9%   ↓ 5%   ↓ 6%   ↓ 4%   ↓ 12%   plateau
```

Recommendation: Log-transform n_gxs to capture diminishing marginal impact.

---

6. Therapeutic Area Analysis

6.1 Erosion by Therapeutic Area (Ranked)

Rank	Therapeutic Area	Mean Vol Norm	Erosion Level
1	Anti-infectives	0.515	Highest
2	Antineoplastic & Immunology	0.551	High
3	Musculoskeletal/Rheumatology	0.557	High
4	Parasitology	0.591	Medium
5	Cardiovascular/Metabolic	0.592	Medium
6	Haematology	0.597	Medium
7	Nervous System	0.607	Medium
8	Obstetrics/Gynaecology	0.609	Medium
9	Systemic Hormones	0.628	Medium
10	Others	0.630	Medium
11	Dermatology	0.644	Low
12	Endocrinology/Metabolic	0.665	Low
13	Respiratory/Immuno-inflammatory	0.698	Low
14	Sensory Organs	0.725	Lowest

6.2 Key Therapeutic Area Insights

High Erosion Areas (Mean < 0.56):

- Anti-infectives: Highly substitutable antibiotics
- Oncology/Immunology: High-cost specialty drugs with biosimilar pressure
- Musculoskeletal: Generic-friendly pain medications

Low Erosion Areas (Mean > 0.65):

- Sensory Organs: Specialized formulations, patient loyalty
- Respiratory: Complex delivery systems (inhalers)

- Endocrinology: Chronic conditions, switching costs

Range: 0.515 to 0.725 (21 percentage points spread)

---

7. Biological vs Small Molecule

7.1 Comparison Summary

Drug Type	Mean Vol Norm	Final Vol (M23)	Total Erosion
Small Molecule	0.593	0.487	42.57%
Biological	0.619	0.488	42.87%

7.2 Key Finding

Surprisingly similar erosion patterns!

- Both drug types show nearly identical total erosion (~42.7%)
- Biological drugs have slightly higher mean during erosion period
- Final equilibrium is virtually the same (0.487 vs 0.488)

Possible Explanations:

1. Biosimilars have become competitive with generics
2. Dataset may include newer biosimilar-exposed biologics
3. Price pressure affects both categories similarly

Recommendation: `biological` feature has **limited predictive power** for final erosion level.

---

8. Hospital Rate Impact

8.1 Erosion by Hospital Distribution

Hospital Rate	Month 0	Month 12	Month 23	Pattern
0-25% (Retail)	0.874	0.584	0.509	Moderate erosion
25-50%	0.871	0.627	0.556	Slowest erosion
50-75%	0.853	0.627	0.485	Moderate erosion
75-100% (Hospital)	0.849	0.554	0.446	**Fastest erosion**

8.2 Key Insights

High Hospital Rate (75-100%) Shows:

- Faster initial erosion (hospital tender processes)
- Lower final equilibrium (0.446 vs 0.509-0.556)
- More price-sensitive purchasing decisions

Low Hospital Rate (0-25%) Shows:

- Slower erosion trajectory
- Higher final volume retention
- Patient/prescriber loyalty matters more

Recommendation: Create hospital_rate_bucket feature for better segmentation.

---

9. Erosion Speed Metrics

9.1 Time to 50% Volume

Bucket	Mean (months)	Median (months)	Min	Max
Bucket 1	1.37	1.0	0	6
Bucket 2	7.92	6.5	0	23

Bucket 1 reaches 50% erosion 5.8× faster than Bucket 2

9.2 First 6 Months Erosion Rate

Bucket	Mean Erosion (6m)	Median
Bucket 1	66.7%	68.3%
Bucket 2	24.3%	21.1%

Bucket 1 loses 2.7× more volume in first 6 months

9.3 Final Equilibrium (Months 18-23)

Bucket	Mean Final Vol	Median
Bucket 1	0.077	0.063
Bucket 2	0.529	0.539

Bucket 1 final level is 6.9× lower than Bucket 2

---

10. Key Patterns & Insights

10.1 Temporal Patterns

```
EROSION TIMELINE
================
Month 0-3:   Rapid initial drop (most critical period)
Month 3-6:   Continued steep decline
Month 6-12:  Erosion slows, pattern emerges
Month 12-18: Approaching equilibrium
```

```
Month 18-23: Stable equilibrium level

🎯 CRITICAL INSIGHT: First 6 months determine bucket membership!
```

10.2 Feature Importance Hierarchy (Expected)

Based on EDA, expected feature importance:

1. `months_postgx` - Time is primary driver
2. `n_gxs` - Competition pressure (diminishing returns)
3. `ther_area` - Strong category effects
4. `hospital_rate` - Distribution channel impact
5. `avg_vol` / `avg_j` - Baseline volume matters
6. `biological` - Limited impact on final erosion

10.3 Bucket Prediction Signals

Early Bucket 1 Signals:

- Sharp initial drop (>40% in month 0)
- Rapid competition buildup
- Anti-infectives or Oncology therapeutic area
- High hospital rate (>75%)

Bucket 2 Signals:

- Gradual initial decline (<20% in month 0)
- Lower competition
- Sensory organs, Respiratory areas
- Lower hospital rate

---

11. Data Preprocessing Recommendations

11.1 Missing Value Handling

Issue	Recommended Action
`time_to_50pct` is NaN for brands never reaching 50%	Impute with 24 (max months) or create binary flag `reached_50pct`
Missing `avg_vol` for some test brands	Use training median or predict from features

11.2 Outlier Treatment

Variable	Outlier Pattern	Recommendation
`n_gxs`	Few brands with 30-43 competitors	Cap at 15 (99th percentile)
`vol_norm`	Some values > 1.0 (growth cases)	Keep, but flag for analysis

Variable	Outlier Pattern	Recommendation
`volume`	High variance across brands	Use normalized volume instead

11.3 Feature Scaling

```
# Recommended scaling strategy
scaling_strategy = {
    'months_postgx': 'StandardScaler',    # Already bounded, center important
    'n_gxs': 'Log1p + StandardScaler',    # Diminishing returns pattern
    'hospital_rate': 'None or MinMax',    # Already 0-1 range
    'avg_vol': 'Log + StandardScaler',    # High variance
    'vol_norm': 'None',                   # Target-related, keep as-is
}
```

11.4 Categorical Encoding

Variable	Recommended Encoding
`country`	Target encoding (mean vol_norm)
`ther_area`	Target encoding or ordinal (by erosion rank)
`biological`	Keep binary (0/1)
`main_package`	Target encoding

---

12. Feature Engineering Recommendations

12.1 Time-Based Features

```
# Essential time features
time_features = {
    'months_postgx_squared': 'months_postgx ** 2',  # Capture non-linear decay
    'months_postgx_sqrt': 'sqrt(months_postgx)',    # Early period emphasis
    'is_early_period': 'months_postgx <= 6',        # Critical first 6 months
    'is_late_period': 'months_postgx >= 18',        # Equilibrium period
    'time_bucket': 'pd.cut(months_postgx, [0,6,12,18,24])',  # Time segments
}
```

12.2 Competition Features

```
# Competition-based features
competition_features = {
```

```
    'log_n_gxs': 'np.log1p(n_gxs)',                # Diminishing returns
    'n_gxs_squared': 'n_gxs ** 2',                 # Non-linear impact
    'has_competition': 'n_gxs > 0',                # Binary flag
    'high_competition': 'n_gxs >= 5',              # Threshold flag
    'competition_intensity': 'n_gxs / (months_postgx + 1)',  # Rate of buildup
}
```

12.3 Lag Features (Critical for Time Series)

```
# Lag features for volume trajectory
lag_features = {
    'vol_norm_lag1': 'Previous month volume',
    'vol_norm_lag3': '3-month lag',
    'vol_norm_lag6': '6-month lag',
    'vol_norm_diff1': 'Month-over-month change',
    'vol_norm_diff3': '3-month change',
    'vol_norm_pct_change': 'Percentage change',
}
```

12.4 Rolling Statistics

```
# Rolling window features
rolling_features = {
    'vol_norm_rolling_mean_3': '3-month rolling average',
    'vol_norm_rolling_mean_6': '6-month rolling average',
    'vol_norm_rolling_std_3': '3-month rolling std',
    'vol_norm_rolling_min_6': '6-month rolling minimum',
    'erosion_rate_3m': 'Erosion rate over last 3 months',
}
```

12.5 Interaction Features

```
# High-value interaction features
interaction_features = {
    'time_x_competition': 'months_postgx * n_gxs',
    'time_x_hospital': 'months_postgx * hospital_rate',
    'competition_x_hospital': 'n_gxs * hospital_rate',
    'early_high_competition': 'is_early_period * high_competition',
}
```

12.6 Category-Based Features

```
# Category-derived features
category_features = {
    'ther_area_erosion_rank': 'Ordinal ranking by mean erosion',
    'ther_area_mean_erosion': 'Target-encoded therapeutic area',
    'country_mean_erosion': 'Target-encoded country',
    'is_high_erosion_area': 'ther_area in [Anti-infectives, Oncology, MSK]',
}
```

---

13. Modeling Recommendations

13.1 Handling Class Imbalance

Critical: Bucket 1 is only 6.7% of data but has 2× weight!

Recommended Approaches:

1. **Stratified Sampling**

```
# Use stratified K-fold
from sklearn.model_selection import StratifiedKFold
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
```

2. **Sample Weights**

```
# Weight Bucket 1 samples higher
sample_weight = df['bucket'].map({1: 2.0, 2: 1.0})
model.fit(X, y, sample_weight=sample_weight)
```

3. **Oversampling Bucket 1**

```
# SMOTE for Bucket 1 brands
from imblearn.over_sampling import SMOTE
X_res, y_res = SMOTE().fit_resample(X, y_bucket)
```

4. **Separate Models Per Bucket**

```
# Train bucket classifier first, then separate regressors
bucket_classifier.fit(X, y_bucket)
bucket1_model.fit(X[bucket==1], y[bucket==1])
bucket2_model.fit(X[bucket==2], y[bucket==2])
```

13.2 Model Selection

Model Type	Strengths	Use Case
Exponential Decay	Captures physics of erosion	Baseline, interpretable
LightGBM/XGBoost	Handles non-linearity, interactions	Primary ML model
Hybrid (Physics+ML)	Physics baseline + ML residuals	Best of both worlds
ARIHOW (SARIMAX+HW)	Time series patterns	Per-brand forecasting

13.3 Cross-Validation Strategy

```
# Recommended CV approach
cv_strategy = {
    'type': 'GroupKFold',           # Group by brand to prevent leakage
    'n_splits': 5,
    'stratify_by': 'bucket',        # Ensure both buckets in each fold
    'time_aware': True,             # Don't use future to predict past
}
```

13.4 Evaluation Metrics

Metric	Purpose	Weight
WMAPE (Scenario metric)	Primary competition metric	100%
MAE by Bucket	Ensure Bucket 1 not ignored	Diagnostic
MAE by Time Period	Check early vs late prediction quality	Diagnostic

13.5 Ensemble Strategy

```
# Recommended ensemble
ensemble = {
    'baseline_exp_decay': 0.3,      # Physics-based foundation
    'hybrid_lightgbm': 0.4,         # Primary ML model
    'arihow': 0.3,                  # Time series component
}
```

13.6 Post-Processing

1. **Clip Predictions**: Ensure vol_norm ∈ [0, 1.5]
2. **Monotonicity**: Enforce non-increasing volume (optional)
3. **Bucket-Specific Adjustments**: Scale predictions if bucket predictions systematically biased

---

Summary: Top 10 Action Items

Priority	Action	Impact
1	Handle Bucket 1 imbalance (sample weights or separate models)	Critical
2	Create lag features (vol_norm_lag1, lag3, lag6)	High
3	Add rolling statistics (3m, 6m windows)	High
4	Log-transform n_gxs to capture diminishing returns	Medium
5	Target-encode therapeutic area by erosion rank	Medium
6	Create time buckets (0-6, 6-12, 12-18, 18-24)	Medium
7	Add erosion rate features (month-over-month change)	Medium
8	Cap n_gxs at 15 (outlier handling)	Low
9	Create hospital_rate buckets (4 levels)	Low
10	Add interaction features (time × competition)	Low

Appendix: Data Files

File	Description
`fig01_bucket_distribution.json/csv`	Bucket counts and percentages
`fig02_erosion_curves.json/csv`	Monthly vol_norm by bucket
`fig03_sample_trajectories.json/csv`	Individual brand examples
`fig04_competition_impact.json`	n_gxs vs vol_norm analysis
`fig04_n_gxs_impact.csv`	Detailed competition data
`fig04_competition_trajectory.csv`	n_gxs over time
`fig05_therapeutic_areas.json/csv`	Erosion by ther_area
`fig06_biological_vs_small.json/csv`	Drug type comparison
`fig07_hospital_rate.json/csv`	Hospital channel impact
`fig08_erosion_speed.json`	Speed metrics summary

File	Description
`fig08_erosion_speed_full.csv`	Per-brand erosion metrics
`eda_complete_summary.json`	Full EDA summary

Report generated from EDA data files in `reports/eda_data/`