NOVARTIS 2025

Novartis Datathon 2025

Generic Erosion Forecasting

Forecasting post–Loss of Exclusivity volume erosion of branded drugs


Time Series


Data Engineering


Machine Learning


Forecasting

Data Science | Business Analytics

Data Sources & Feature Engineering

Comprehensive data integration and scenario-aware feature engineering

Data Sources

df_volume_*
Monthly volumes, months_postgx, per (country, brand_name)

df_generics_*
Number of generic competitors (n_gxs) over time

df_medicine_info_*
Drug characteristics (ther_area, hospital_rate, biological, small_molecule, main_package)

Panel Construction

Key: (country, brand_name, months_postgx)

Derived: avg_vol_12m, mean_erosion, bucket (1 or 2)

Feature Engineering

Pre-entry Statistics

- Rolling averages (3/6/12 months)
- Volatility and trend features

Time & Seasonality

- months_postgx, months_postgx²
- Time buckets, calendar features

Generics Dynamics

- n_gxs, has_generic
- cummax_n_gxs, time to first generic

Drug Attributes

- Therapeutic area
- Hospital vs retail, biological vs small molecule

Leakage Prevention

- Never use bucket, mean_erosion, y_norm, or volume as features
- country and brand_name treated as meta only
- Scenario 2 only: Early-signal features over months 0–5

Modeling Approach & Pipeline

Unified scenario-aware pipeline with CatBoost as the hero model

Models Explored

Tree-based / Tabular

CatBoost (hero model)

Classical Baselines

Linear models, HistoricalCurve, Simple baselines

Time-series / Hybrid

ARIHOW, Hybrid physics + ML

Neural Models

MLP, LSTM, CNN-LSTM, KG-GCN-LSTM

Training Strategy

Scenario-specific Training

S1: Train on pre-LOE history, predict months 0–23

S2: Train on history up to month 5, predict months 6–23

Validation Strategy

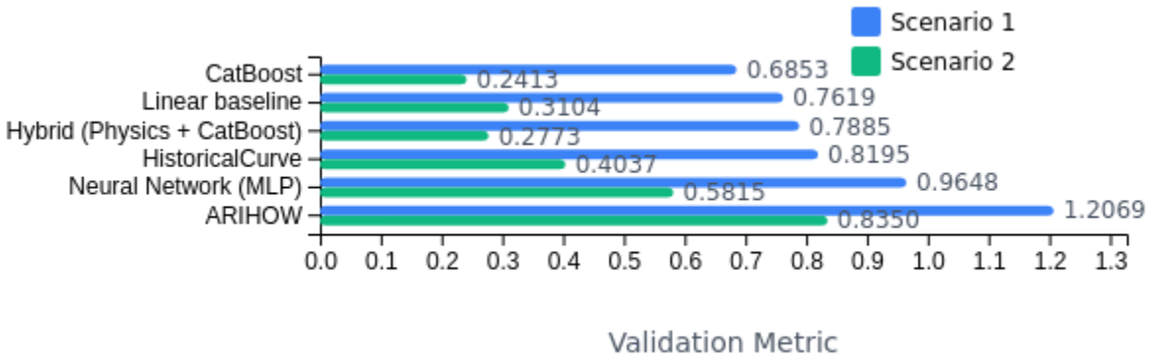
- 80/20 split at series level
- Stratified by bucket
- Early stopping on validation metric

Model Performance & Results

CatBoost outperforms all baselines with optimized hyperparameters

Model Comparison

Scenario 1 | **Scenario 2**



CatBoost dominates across both scenarios

Hero CatBoost Configuration

Depth	Learning Rate
6	0.03

L2 Regularization	Early Stopping
3.0	100 rounds

Hero Run Metrics

Scenario 1

Metric 1	RMSE	MAE
0.7692	0.2488	0.1795

Scenario 2

Metric 2	RMSE	MAE
0.2742	0.2055	0.1265

Business Impact & Future Roadmap

Model enables demand forecasting and pricing strategies while providing framework for future enhancements

Business Impact

Demand Forecasting

Predicts erosion curves 24 months after LOE at country/brand level

Pricing Strategies

Supports launch and pricing strategies for originator and generic manufacturers

Risk Identification

Identifies high-risk high-erosion series early, allowing proactive management

Future Roadmap

Robust Ensembles

Explore ensembles combining CatBoost, Hybrid models, and HistoricalCurve

Additional Model Comparison

Run LightGBM/XGBoost on Linux/GPU to compare against CatBoost

External Signals

Incorporate macroeconomic indicators and competitor launches

Advanced Sequence Models

Explore CNN-LSTM and transformer models after robust baselines are exploited