12Go Report

Results

By applying stepwise feature selection using CatBoost across stratified transport modes, we uncovered significant predictive heterogeneity that was previously masked by the aggregate analysis. While the global model achieved a Lift of 2.73, the mode-specific analysis revealed that the Flight segment is highly deterministic, achieving a **Lift of 14.73** in the top decile. Notably, the Taxi segment demonstrated strictly linear behavior, with the Linear Baseline outperforming the CatBoost model (Lift 6.52 vs 4.94); consequently, adding non-linear complexity yielded diminishing returns for this specific utility-driven mode. Ground transport segments (Bus, Van, Ferry) maintained consistent non-linear lifts ranging from 2.88 to 4.56. All tests are repeatable given random seed 42.

Data cleaning

Filtration and Leakage Control

- Features with 95%+ sparsity were later discarded.
- zero-variance `saw_price` column was excised.
- `price_paid` was zero-filled for auditing but excluded from training to prevent look-ahead bias.

Normalization and Structural Recovery

- `visit_date` was normalized. Missing `trip_id` identifiers were deterministically reconstructed via a composite key (origin, destination, mode).
- To resolve sparsity in `ATC_rate_last_24h` (2.00% missing), values were approximated using intra-day medians grouped by `trip_id` and `visit_date`.
- For `competitor_price`, a Route-to-Price lookup map was utilized to impute missing entries via `trip_id` mapping.
- Logical backfilling was applied to `clicked_trip`: records indicating a booking or cart addition were force-assigned a value of 1, with residual nulls zero-filled.

Validated Conditional Imputation

`price_shown` (1.98% missing) was reconstructed using logic-based grouping. Hypothesis testing confirmed significantly lower price variance within daily windows (+/-$30.91) versus monthly (+/-$36.63) (For experimental group). Consequently, nulls were resolved by grouping by Visit Day and ticket attributes:

- **Unflagged:** Reverted to baseline.
- **Flagged:** Imputed via daily markup/discount sub-cluster means to preserve bimodal distribution.

Exploratory Data Analysis (EDA)

Conversion Funnel Dynamics

Analysis of the aggregate user journey (N=5,000) reveals significant drop-off post-engagement, with a Click-Through Rate (CTR) of 24.7% and an ultimate Booking Conversion Rate (CVR) of 4.9%.

Revenue Integrity & Paradox Mitigation

While the Dynamic group yielded a 31% increase in total revenue, the magnitude of this gain necessitated a rigorous search for Simpson's Paradox or covariate imbalance. The following dimensions were audited and found to be statistically homogenous across both groups:

- **Trip Logistics:** No significant variance in mean distance, travel duration, or transport type distribution.
- **Historical Performance:** Baseline popularity scores and 7-day conversion rates were uniform.
- **User Attribution:** No skew detected in device type, OS, UTM source, or geographic origin.

The absence of confounding variables suggests that the observed revenue lift is attributable to the pricing strategy rather than sample bias.

Feature Engineering

Following imputation, two synthetic features were engineered to capture relative value and price density:

- **Price Density** (price_per_km)**:** Calculated as price_shown/(distance_km + 1) to normalize cost across varying trip lengths while preventing division by zero.
- **Composite Value Score (**valuable_deal**):** An interaction term multiplying price_ratio_vs_competitor by popularity_score. This feature quantifies the deal strength by weighting competitive pricing against historical demand.

Model Selection & Experimental Framework

The predictive pipeline employed a multi-stage benchmarking framework evaluating XGBoost, CatBoost, Random Forest, TabNet, and Balanced Bagging against a linear baseline via a stratified training, testing, and validation protocol. Look-ahead leakage from technical features was systematically mitigated to ensure model generalizability. While the initial aggregate analysis yielded competitive results with TabNet, subsequent error analysis revealed that global modeling masked distinct behavioral signals across transport modes. Consequently, the framework was restructured to deploy **CatBoost Stepwise Selection** within stratified transport segments.

CatBoost was selected for this granular phase due to its superior handling of categorical distinctions and stability on smaller, mode-specific subsets. To address the high-dimensional sparsity resulting from one-hot encoding, a stepwise feature selection was executed for each transport mode. This localized refinement pruned non-contributing dimensions, allowing the model to capture specific pricing sensitivities—such as those in the Flight segment—that were previously diluted in the global model. This pivot from a global deep learning architecture to mode-specific gradient boosting was validated by the localized performance lifts, most notably the 14.73x Lift achieved in the Flight segment.

The Simulation

While the initial aggregate model suggested market-wide price inelasticity (0.00), the granular CatBoost Stepwise simulation revealed critical behavioral heterogeneity that aggregation had masked. Most notably, the Van segment exhibited high price sensitivity with an elasticity of -0.758, indicating that a 10% price increase triggers a ~7.6% drop in volume. Conversely, Bus and Ferry segments confirmed true inelasticity (~0.00), while Flights showed only marginal sensitivity (-0.052).

This dichotomy invalidates a uniform pricing strategy. The simulation proves that yield management must be applied differentially: aggressive markups are safe for inelastic modes (Bus/Ferry), but a nuanced approach is required for Vans to prevent revenue-destroying churn.