12Go Report

Results

By applying stepwise feature selection to a TabNet model, we achieved a Lift of 2.73 in the top decile, confirming that the selected features are strong drivers of the target outcome. Notably, the analysis revealed that the target is price inelastic; consequently, adding pricing features yielded no improvement in predictive power. 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. TabNet (Deep Learning) was identified as the optimal architecture for the booking stage in terms of PR-AUC, while CatBoost demonstrated superior probability calibration as measured by the Brier Score. The final TabNet iteration achieved a Lift of 1.77 in the top decile, demonstrating significantly higher precision in isolating high-conversion segments compared to other methods.

To address the high-dimensional sparsity resulting from one-hot encoding, a bidirectional stepwise selection was executed on the TabNet booking target. Given TabNet's native sequential attention, the procedure prioritized forward selection to isolate potent feature interactions within the sparse matrix. This refinement pruned non-contributing dimensions, increasing test-set Lift from 1.77 to 2.73. This 54% performance increase confirms the architecture's capacity to distill latent intent signals from sparse feature spaces, effectively bypassing price-inelastic noise.

The Simulation

To evaluate the economic viability of the predictive models, we implemented a Counterfactual Simulation designed to measure the trade-off between profit margins and customer attrition. This engine interrogates each model by presenting it with identical user sessions under varying price conditions to observe the resulting shift in Predicted Probability.