# Feature Engineering Summary (Final)

## Features Used in Final Model

### price\_rank

* • Creation: Rank of `price\_usd` within the same `srch\_id` (1 = cheapest).
* • Meaning: Captures how cheap the hotel is relative to others shown to the user.

### star\_rank\_pct

* • Creation: Percentile rank of `prop\_starrating` within the same search (higher is better).
* • Meaning: Provides a relative quality score that is comparable across searches.

### value\_score

* • Creation: `prop\_review\_score` / (`price\_usd` + 1e‑3).
* • Meaning: Simple ‘bang‑for‑buck’ indicator.

### price\_rel\_search\_rank

* • Creation: Percentile rank of `price\_usd / median\_price(search)` in the search.
* • Meaning: Smooth 0‑to‑1 version of relative price.

### star\_diff\_search

* • Creation: `prop\_starrating` − mean star rating in the search.
* • Meaning: Positive if the hotel is better than the page average.

### review\_diff\_search

* • Creation: `prop\_review\_score` − mean review score in the search.
* • Meaning: Shows above‑ or below‑average customer feedback.

### residual\_price

* • Creation: `price\_usd` − mean `price\_usd` for the same `prop\_id`.
* • Meaning: Price deviation from a hotel’s typical rate; flags temporary discounts/premiums.

### prop\_popularity

* • Creation: Smoothed booking rate per property: (bookings + 10 × global\_rate) / (impressions + 10).
* • Meaning: Overall demand indicator; high‑popularity properties convert better.

### dest\_popularity

* • Creation: Smoothed booking rate per `srch\_destination\_id`.
* • Meaning: Captures destination‑level demand swings (events, seasonality).

### location\_score\_gap

* • Creation: `prop\_location\_score1` − mean location\_score1 in the search.
* • Meaning: Relative desirability of the hotel’s location.

### location\_score2\_gap

* • Creation: `prop\_location\_score2` − mean location\_score2 in the search.
* • Meaning: Second dimension of relative location quality.

### cheap\_and\_good

* • Creation: `price\_rank` × `star\_diff\_search`.
* • Meaning: Highlights listings that are inexpensive \*and\* above‑average quality.

### search\_size

* • Creation: Count of hotels shown in the same `srch\_id`.
* • Meaning: Proxy for supply / competition intensity in the search results.

### price\_per\_night

* • Creation: `price\_usd / srch\_length\_of\_stay`.
* • Meaning: Normalises price across different stay lengths.

### price\_per\_night\_rel

* • Creation: `price\_per\_night / median(price\_per\_night) in search`.
* • Meaning: Relative nightly cost.

## Features Tested and Removed

### stay\_has\_weekend

* • Creation: Flag for stays overlapping a weekend (original version).
* • Reason for removal: Removed: overlapped with new arrival\_weekend; SHAP ≈ 0.

### price\_rel\_search

* • Creation: `price\_usd / median\_price(search)` – raw ratio.
* • Reason for removal: Removed: rank variant captured the signal better; raw ratio redundant.

### price\_discount\_pct

* • Creation: Discount vs property’s historical average price.
* • Reason for removal: Removed: sparse (needs historical price); marginal SHAP impact.

### distance\_pct

* • Creation: Percentile rank of distance to destination within search.
* • Reason for removal: Removed: data too noisy / many missing distances.

### pref\_star\_gap / pref\_price\_gap

* • Creation: Difference between hotel stars / price and user’s historical preference.
* • Reason for removal: Removed: user history mostly missing; negligible lift.

### competitor aggregates (comp\_rate\_mean, comp\_any\_cheaper, comp\_best\_rate, comp\_any\_no\_inv)

* • Creation: Various summaries of competitor price & availability fields.
* • Reason for removal: Removed: sparse & outdated competitor data → near‑zero SHAP.

### booking\_window\_bin

* • Creation: Binned `srch\_booking\_window`.
* • Reason for removal: Removed: replaced by smoother cyclical check‑in features.

### early\_price\_rank

* • Creation: `price\_rank × (booking\_window > 30)`.
* • Reason for removal: Removed: interaction added little once residual\_price present.

### comp\_price\_gap

* • Creation: Min competitor percentage difference (dense numeric).
* • Reason for removal: Removed: also negligible once residual\_price & popularity features used.

### month\_srch

* • Creation: Raw month of search (1‑12).
* • Reason for removal: Removed: cyclical encode of check‑in month worked better.

### log\_dest\_distance

* • Creation: log1p of `orig\_destination\_distance`.
* • Reason for removal: Removed: many NAs; distance not predictive in this data.

### rooms\_per\_guest

* • Creation: `srch\_room\_count / group\_size`.
* • Reason for removal: Removed: low SHAP; didn’t add unique signal.

### checkin\_dow\_sin

* • Creation: Sine of check-in day-of-week (cyclical).
* • Reason for removal: Removed: negligible SHAP; pattern captured elsewhere.

### checkin\_dow

* • Creation: Day-of-week integer (0=Mon).
* • Reason for removal: Removed: captured already via other time signals.

### group\_size

* • Creation: `srch\_adults\_count + srch\_children\_count`.
* • Reason for removal: Removed: not predictive enough in this data.

### has\_history

* • Creation: Whether user has booking history (1/0).
* • Reason for removal: Removed: most users lack history; adds noise.

### checkin\_dow\_cos

* • Creation: Cosine of check-in day-of-week.
* • Reason for removal: Removed: same reason as `checkin\_dow\_sin`.

### price\_per\_night\_rel\_rank

* • Creation: Percentile rank of relative nightly price.
* • Reason for removal: Removed: overlap with price rank & price\_per\_night.

### children\_bool

* • Creation: 1 if children included, else 0.
* • Reason for removal: Removed: low impact, even on family-targeted searches.

### length\_of\_stay\_bin

* • Creation: Binned stay duration.
* • Reason for removal: Removed: adds little beyond raw length or nightly price.

### arrival\_weekend

* • Creation: 1 if arrival is on Fri/Sat.
* • Reason for removal: Removed: insignificant SHAP impact.

### is\_chain

* • Creation: 1 if part of a hotel chain.
* • Reason for removal: Removed: SHAP near zero.

### chain\_popularity

* • Creation: Booking rate of chains vs independents.
* • Reason for removal: Removed: not adding lift beyond popularity and rating.