Boosting Versus Foundation Models for Ordinal Home-Sharing Engagement Classification in Latin America

Pedro Shiguihara[1], Nils Murrugarra[2]
Department of Information Systems Engineering,
Universidad San Ignacio de Loyola, Lima, Peru[1]
Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, USA[2]

***Abstract*—Review activity is the standard public proxy of guest engagement on peer-to-peer accommodation platforms, yet is modeled almost exclusively as count regression. This study presents a three-level engagement classification of peer-to-peer home-sharing in Latin America: 76,624 listings from Buenos Aires, Santiago, and Mexico City, with tercile labels from training folds only. A pre-specified, amendment-logged evidence ladder of fifteen attempts plus three contingency campaigns measures which interventions affect macro-averaged F1. A tuned LightGBM over engineered features reaches average macro F1 0.8249, 23.0 points above the best cold-start result (0.5945); a feature-level temporal audit lowers the deployable cold-start reference to 0.5301, widening the gap to 29.5 points. Three findings emerge. First, every full-scenario result lands in a narrow band (0.81–0.83) that doubled budgets do not move: an observed performance plateau within the evaluated feature and model space. Second, Spanish text helps in the cold-start scenario (+2.4 points nominally, +5.3 over the audited strict set) but adds little once review-derived features are present, and sparse lexical vectors match dense multilingual E5 embeddings at far lower cost. Third, an untuned tabular foundation model reaches 0.8205 under a pre-specified 10,000-row context subsample. A clearly labeled post-hoc re-evaluation at full context, outside the pre-specified protocol, reaches 0.8266, the best average observed, which an exploratory multi-resampling confirmation upholds in all twelve paired comparisons. Ordinal metrics, per-class breakdowns, and a Frank–Hall baseline locate the residual error in the middle tercile; repeated outer resampling bounds selection optimism at 0.2–0.3 points.**

Keywords*—Peer-to-peer home-sharing; ordinal classification; engagement prediction; gradient boosting; multilingual text embeddings; hyperparameter optimization; tabular foundation models; TabPFN; Latin America

I. Introduction

Peer-to-peer (P2P) accommodation platforms publish, through initiatives such as Inside Airbnb [1], one of the few large-scale open records of digital marketplace activity. Within these records, the number of reviews a listing accumulates is the standard observable proxy of guest engagement and, indirectly, of demand: reviews are strongly coupled to completed stays, are visible to all market participants, and drive both ranking and trust formation [2], [3], [4]. A large econometric literature has therefore modeled review counts with Poisson, negative binomial, and hurdle regressions across dozens of cities [5], [6], [7], [8], while the machine learning literature on these platforms has concentrated overwhelmingly on price prediction [9], [10], [11], [12].

Between these two bodies of work lies a practical gap. Hosts, platform operators, and municipal regulators rarely need the exact expected count of future reviews; they need to know whether a listing will attract low, medium, or high engagement. Casting the problem as classification makes the output directly actionable and enables evaluation with class-sensitive metrics such as macro-averaged F1, which weight the scarce high-engagement and zero-engagement listings as heavily as the bulk of the market. Yet, to the best of our knowledge, no published study reports macro F1, precision, or recall for engagement classification on Inside Airbnb data from Buenos Aires, Santiago de Chile, or Mexico City. The closest antecedent we identified, Kirkos [13], classifies listings of a single European city (Thessaloniki) into two classes split at the mean of the engagement measure, a threshold that collapses the heavy right tail typical of review-count distributions. Subroto and Christianis [14] classify review ratings, a different target, and report accuracy rather than macro-averaged metrics.

This study addresses that gap with a three-level engagement classification of 76,624 listings across the three cities. The contributions are twofold:

- Three-level engagement classification for Latin American home-sharing data. A distribution-aware target with ordered labels built from per-fold training terciles over 76,624 listings of Buenos Aires, Santiago, and Mexico City, evaluated with macro F1, precision, and recall under stratified five-fold cross-validation, complemented post hoc by per-class breakdowns, ordinal metrics, and a Frank–Hall ordinal baseline.
- A quantified assessment of which interventions affect performance. A pre-specified, amendment-logged evidence ladder of 15 executed attempts plus three contingency campaigns (27,219 registered trials in 30 Optuna studies, of which 12,552 completed the full five-fold evaluation) measures the marginal value of Spanish text (sparse term frequency–inverse document frequency (TF-IDF) vectors versus multilingual E5 embeddings), engineered features, out-of-fold stacking, calibration, a zero-hurdle design, doubled search budgets, and a tabular foundation model [15] evaluated as a domain study in short-term rentals. The ladder

converges to a stable, observed performance plateau at macro F1 0.81–0.83 within the evaluated feature and model space.

The study is organized around three research questions.

RQ1: what is the marginal value of the Spanish description text, sparse versus dense (Section VII-C)?

RQ2: which contributes more, model-centric sophistication or feature engineering (Section VII-D)?

RQ3: can a tuning-free tabular foundation model match tuned gradient boosting at this scale (Sections VII-E and VII-F)?

This article proceeds as follows. Section II reviews related work, and Section III introduces the concepts and model families the study builds on. Section IV describes the data and the problem formulation. Section V details the methodology and Section VI the experimental infrastructure. Section VII reports results, Section VIII discusses them, and Section IX concludes with limitations and future work.

II. Related Work

Review counts on P2P platforms are heavily overdispersed and zero-inflated, which has made count regression the dominant tool. Biswas et al. [5] combine clustering with Poisson-family regressions over home-sharing listings; Sengupta et al. [6] use hurdle models over cities in four continents; Chen et al. [7] link host online affinities to review volume; and Zhan and Cheng [8] take a multimodal view of host identity. Methodologically these works build on the classical count-regression corpus [16], [17], and they assess fit on the count scale rather than with class-sensitive classification metrics.

Classification studies are scarce. The closest peer-reviewed antecedent, Kirkos [13], is a binary high/low split at the mean for Thessaloniki, with correlation-based feature selection, the synthetic minority oversampling technique (SMOTE), and 10-fold cross-validation, reaching F-measure 0.792–0.835 with random forests. Subroto and Christianis [14] predict high versus low review ratings in Indonesia with tree ensembles and neural networks, reporting accuracy up to 84.79%. Neither work is ordinal, multi-city, macro-averaged, nor Spanish-language, and neither evaluates modern gradient boosting with large-scale hyperparameter optimization (HPO).

Gradient-boosted decision trees (GBDTs), namely XGBoost [18], LightGBM [19], and CatBoost [20], remain the reference for medium-sized tabular problems [21], [22], with HPO contributing measurable but concentrated gains [23]. Tabular foundation models have recently challenged this dominance: TabPFN v2 reports superiority over tuned GBDTs on small datasets [15], and TabPFN-2.5 extends the applicable range to 50,000 rows [24]. A rapidly growing wave of 2024–2026 domain studies has since taken these models well beyond generic benchmark suites, into materials informatics [25], industrial-IoT intrusion detection [26], cheminformatics [27], and clinical risk prediction [28]. These studies repeatedly report an untuned foundation model competitive with, and in small-sample or out-of-distribution regimes superior to, tuned gradient boosting and other strong baselines. Home-sharing datasets have appeared in generic tabular benchmarks, but short-term rentals remain absent from this wave of domain studies; we fill that gap. For ensembling, out-of-fold stacking is the mechanism behind AutoGluon's strong tabular results [29].

Adding listing text improves price regression on home-sharing listings [9], fusing structured listing attributes with review text processed by natural language models nowcasts neighbourhood change on the same platform [30], and benchmark studies show that stacking text representations onto tabular features is the most reliable fusion strategy, while also warning that word-level n-gram baselines remain difficult to outperform [31]. Multilingual encoder families such as E5 provide mature embeddings for Spanish [32]. The marginal value of Spanish-language listing text for engagement classification remains unmeasured in this literature; our ladder quantifies it in both a cold-start and a full-information scenario.

The dominant recipe in the P2P classification literature is flat multiclass plus SMOTE [13]. Recent evidence questions synthetic oversampling: for strong learners, class weights plus calibrated thresholds match or beat SMOTE [33], [34]. Ordinal structure can be exploited with cumulative binary decompositions [35], and zero inflation with two-stage hurdle designs [36]. Our ladder evaluates class weighting, calibration with tuned per-class thresholds, and a zero-hurdle stage empirically. The pre-specified attempts treat the task as flat three-class classification; a post-hoc follow-up adds order-sensitive metrics and a Frank–Hall baseline (Section VII-F). We follow Demšar's guidance [37] for paired comparisons.

III. Theoretical Background

This section introduces the concepts the rest of the study relies on: the construction of an ordinal target from counts, macro-averaged evaluation, gradient-boosted decision trees, large-scale hyperparameter optimization, tabular foundation models, and the text representations fused with the tabular features.

A. From Counts to an Ordinal Target

Review counts are non-negative, heavily overdispersed, and zero-inflated, which is why they are classically modeled with Poisson and negative binomial regression [16], [17]. An alternative is to discretize the count into K ordered levels and predict the level directly. When the cut points are quantiles of the training distribution (here terciles, $K = 3$), the resulting classes are near-balanced by construction and adapt to each market's distribution, unlike a split at the mean, which the heavy right tail pulls upward. The levels retain a natural order (low $\prec$ medium $\prec$ high); ordinal-aware learners can exploit that order, for example through cumulative binary decompositions [35], although the problem can also be addressed with standard (flat) multiclass classifiers, the route taken by the executed attempts in this study. Zero inflation can additionally be handled with two-stage hurdle designs that first separate zero from non-zero outcomes [36].

B. Macro-Averaged Evaluation

For a class c, precision P_c is the fraction of listings predicted as c that truly belong to c, recall R_c is the fraction of listings of class c that the model recovers, and the per-class

F1 is their harmonic mean. The macro-averaged F1 over K classes,

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{K}\sum_{c=1}^{K}\frac{2\,P_c R_c}{P_c + R_c}, \tag{1}$$

weights every class equally regardless of its support, so scarce classes count as much as the majority class, which is the property that motivates its use for engagement prediction, where the extreme classes matter most. All scores are estimated with stratified k-fold cross-validation, which partitions the data into k folds, holds each fold out once for evaluation while fitting on the remaining $k-1$, and keeps every fold's class proportions close to the full sample's. The resulting intervals estimate the error of the fitting procedure rather than of one fitted model [38], and paired nonparametric tests such as the Wilcoxon signed-rank test compare two learners fold by fold [37].

C. Gradient-Boosted Decision Trees

GBDTs build an additive ensemble of shallow decision trees, each new tree fit to the gradient of the loss of the ensemble built so far. XGBoost introduced regularized boosting with second-order updates [18]; LightGBM accelerates training with gradient-based sampling and leaf-wise tree growth [19]; CatBoost adds ordered boosting and native categorical treatment [20]. On medium-sized tabular data these families remain the reference against deep alternatives [21], [22]. Two standard complements appear in this study: out-of-fold stacking, which trains a meta-learner on cross-validated predictions of several base families [29], and probability calibration followed by per-class decision thresholds.

D. Hyperparameter Optimization

GBDT performance depends measurably on its hyperparameters, with gains that are real but concentrated in a few of them [23]. Sequential model-based methods such as the Tree-structured Parzen Estimator (TPE) model the mapping from configuration to score and concentrate trials in promising regions; Optuna implements TPE with pruning and shared distributed storage [39], and asynchronous schemes such as the Asynchronous Successive Halving Algorithm (ASHA) scale the search to many parallel workers [40].

E. Tabular Foundation Models

A tabular foundation model is a transformer pre-trained on large collections of synthetic tabular prediction tasks. At inference it receives the training rows of a new dataset as context and predicts the labels of test rows in a single forward pass: in-context learning, with no per-dataset training or hyperparameter tuning. TabPFN v2 reports accuracy superior to tuned GBDTs on small datasets [15], and TabPFN-2.5 extends the usable context to about 50,000 rows [24], which makes the approach applicable to corpora of the size studied here.

F. Text Representations for Tabular Fusion

Two standard families turn a free-text field into features. Sparse lexical vectors weight word and word-pair occurrences by TF-IDF, optionally compressed by truncated singular value decomposition (SVD) into a low-dimensional dense representation. Dense semantic embeddings instead map the whole text into a vector space with a pre-trained multilingual encoder such as E5 [32]. Benchmark evidence indicates that concatenating or stacking text representations onto tabular features is the most reliable fusion strategy, and that word-level n-gram baselines remain difficult to outperform [31].

IV. Data and Problem Formulation

A. Corpus

The corpus consists of Inside Airbnb snapshots [1] processed into per-city tables of 19 structured features plus the review-count target, totaling 76,624 listings (Table I).

TABLE I. Corpus Statistics Per City (Review-Count Target)

City	Listings	Mean	Var/mean	Median	Zeros	Max
Buenos Aires	35,172	28.03	72.7	11	16.4%	992
Santiago	15,051	30.19	110.8	9	21.6%	1,227
Mexico City	26,401	52.58	130.1	21	12.8%	1,342
Total	76,624					

The scrape dates are December 27, 2024 for Santiago, January 29–February 2, 2025 for Buenos Aires, and June 25–July 2, 2025 for Mexico City. Each city's table therefore stems from a single scrape window of at most eight days, and the maximum cross-city offset is about six months; Section IV-B analyzes the consequences of this timing. The three markets differ in size and intensity: Mexico City shows the highest median engagement (21 reviews) and the heaviest tail (maximum 1,342), Santiago the highest zero share (21.6%), and Buenos Aires the largest inventory (35,172 listings). Overdispersion is extreme everywhere (variance-to-mean ratios between 72.7 and 130.1), which historically justified negative binomial modeling [17], [16] and, for us, motivates a distribution-aware discretization rather than a mean split.

From the raw snapshots we additionally exploit the Spanish free-text *description* field of each listing to build the text representations of Section V.

B. Three-Level Target

For each city, listings are assigned to three engagement levels with a natural order (low, medium, high) by tercile thresholds of the review count. Two safeguards prevent leakage and preserve comparability. First, the thresholds used to label validation data are always computed on the *training portion of each fold only*. Second, cross-validation strata are fixed with *global* terciles so that fold composition is identical across all experiments. This design warrants clarification: fixing strata with global terciles uses the target's marginal distribution (including test rows) only to balance fold composition (the standard mechanism of stratified cross-validation, where strata are by definition computed on the full sample), while every threshold that produces labels for training or scoring derives

from the training portion of the fold, so no labeling information leaks into evaluation.

Terciles yield near-balanced classes while respecting the heavy tail: unlike a mean split [13], the upper threshold adapts to each market's distribution. Table II reports the global thresholds and the resulting class support: ties at small counts pull the split away from a perfect 33/33/33, with the largest deviation in Santiago (36.3% low class).

TABLE II. GLOBAL TERCILE THRESHOLDS AND CLASS SUPPORT PER CITY

City	Thresholds	Low	Medium	High
Buenos Aires	4 / 25	12,351 (35.1%)	11,344 (32.3%)	11,477 (32.6%)
Santiago	3 / 22	5,462 (36.3%)	4,641 (30.8%)	4,948 (32.9%)
Mexico City	7 / 46	8,906 (33.7%)	8,767 (33.2%)	8,728 (33.1%)

Snapshot timing and comparability. Within each city, all listings stem from one scrape window of at most eight days, so no experiment mixes observation dates: every model is trained and evaluated strictly within a single city's snapshot. The cross-city comparisons in this study are therefore descriptive characterizations of each market at its snapshot moment, never pooled fits. Because review counts are accumulated stocks and the tercile thresholds are recomputed on each fold's training distribution, the class labels are largely invariant to snapshot age. Projecting each listing's count forward by its Inside Airbnb *reviews per month* and relabeling with recomputed terciles changes the class of only 1.7–4.0% of listings under a one-month shift, and of 8.0–13.1% under an extreme six-month shift. This first-order bound ignores listing entry and churn between dates, in this sense, we read the snapshot-timing sensitivity as limited but not zero.

C. The Cold-Start and Full Scenarios

An important circularity affects unrestricted feature sets: several review-derived attributes (for example, review scores or reviews per month) are only observable *after* engagement materializes. We therefore evaluate every idea in two scenarios: *cold-start*, restricted to the 14 structured features whose type of information is, in principle, available when a listing is created (though measured at the snapshot), and *full*, which additionally includes the review-derived features and answers the retrospective question of how well engagement can be explained. Because a snapshot measurement can embed information accumulated after creation, Table III audits each cold feature by its documented mutability; Given the absence of listing-creation timestamps in the available Inside Airbnb data, absolute verification is not feasible, making mutability the most appropriate criterion. Three dynamic, accumulation-prone features (90-day availability, superhost status, host tenure) cannot be verified as creation-time and define, by exclusion, a post-hoc *strict* cold-start variant (Section VII-F). The nominal cold scenario is thus an optimistic upper bound on deployment-time prediction, the strict variant its audited counterpart, and the gap between either and the full scenario quantifies the circularity, itself a finding (Section VII).

D. Leakage Safeguards

Four safeguards protect the evaluation from optimistic bias. First, early stopping is never evaluated on the test fold: each training fold reserves an internal validation split for it. Second, the listing identifier is excluded from every feature matrix. Third, a data audit revealed that the nightly price column, nominally in USD, actually records local currency in all three cities (medians ARS 39,908 in Buenos Aires, CLP 43,857 in Santiago, and MXN 1,039 in Mexico City, with outliers up to 2,632 times the median); the column was renamed and log-transformed per city. Fourth, all label-dependent preprocessing lives inside the per-fold pipeline, while two unsupervised transforms are fit once per city (Section V).

TABLE III. TEMPORAL AUDIT OF THE 14 CREATION-TIME (COLD) FEATURES

Tier	Features	Status
T1: structural, fixed	bedrooms; room type; location exactness	set by the property; retained
T2: host-editable, non-accumulating	summary length; readability; title superlatives; profile photo; amenity count; nightly price; instant bookability; identity verification	editable after creation but do not mechanically accumulate outcome information; retained, with caveat
T3: dynamic, accumulation-prone	90-day availability; superhost status; host tenure (at snapshot)	reflect bookings, review history, or elapsed time; removed in the strict variant

V. METHODOLOGY

A. Overview

Fig. 1 summarizes the experimental design. Rather than a single model comparison, we pre-specified an *evidence ladder*: an ordered sequence of at most 15 attempts (numbered I01–I20 internally; executed attempts keep their original identifiers, so the numbering is not contiguous), each testing one hypothesis, with explicit evidence gates that may skip a planned attempt when an earlier result renders it dominated. Fifteen attempts were executed (I01–I08, I11–I13, I17–I20). I09 was skipped by an evidence gate (Section V-C). I10 and I14–I16, together with a planned Frank–Hall ordinal decomposition, were postponed, not silently discarded, to a second wave by a logged priority amendment when interim evidence placed them off the critical path; the Frank–Hall decomposition was later executed as post-hoc follow-up P3 (Section VII-F). The ladder closed with fourteen of its fifteen slots consumed, because the failed TabPFN run (I18) was re-executed manually without consuming its slot (Section VII-E). The protocol was pre-specified and amendment-logged rather than externally pre-registered. The amendment record is part of the experimental record: soft voting was replaced by out-of-fold stacking for I11, the plan was cut to 15 attempts, and a designed expanded-search phase was canceled. Three contingency campaigns outside the ladder re-explored the most promising preparations with doubled budgets.

B. Feature Preparations

All experiments consume pre-materialized feature matrices ("preparations"), so that every attempt sees exactly the same inputs. Starting from the structured attributes and the Spanish description text, the following transformations define the seven pre-specified preparations evaluated per city; the post-hoc strict variants of Section VII-F are obtained from these by column masking and require no new materialization:

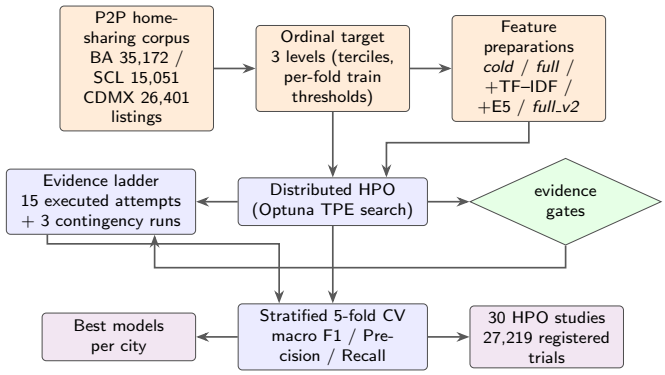

Fig. 1. Experimental pipeline: Three-level target construction, feature preparations, distributed hyperparameter search, evidence-gated attempt ladder, and evaluation.

1) Cold: the 14 creation-time structured features, covering listing structure (bedrooms, room type, amenity count), host attributes (superhost status, identity verification, profile photo, tenure in months), surface signals of the announcement text (summary length, readability, superlatives in the title), location exactness, instant bookability, 90-day availability, and the nightly price, log-transformed per city after the currency audit of Section IV-D;

2) Full: all 19 structured features, adding the five review-derived attributes: the overall review score and four aggregate signals of the received reviews (mean positive and negative sentiment, and the rates of certainty- and cognition-related words); this is the retrospective scenario;

3) Cold_tfidf: / full_tfidf the corresponding structured set concatenated with a sparse lexical representation of the description: each text is stripped of markup, lowercased, and accent-normalized, then vectorized with TF-IDF over word unigrams and bigrams (vocabulary capped at 20,000 terms, discarding terms appearing in fewer than five documents), and the resulting sparse matrix is reduced to its 100 leading components by truncated SVD, so the text contributes 100 dense columns. The TF-IDF/SVD transform is fit once per city on all rows (an unsupervised, label-free fit shared identically by every attempt), not refit inside each training fold; a post-hoc in-fold refit bounds the resulting optimistic bias at +0.1 to +0.4 points of average macro F1, up to +0.7 in a single city (P6, Section VII-F);

4) Cold_e5: / full_e5 the corresponding structured set concatenated with a dense 384-dimensional multilingual-E5-small sentence embedding [32] of the same cleaned description (398 and 403 total columns);

5) Full_v2: the richest preparation (137 columns), extending *full_tfidf* with 18 engineered columns in six groups: G1, amenity counts parsed from the raw amenities list (a total count plus five thematic groups: safety, kitchen, workspace, climate, and comfort); G2, word counts of the host profile and the listing title; G3, three physical-capacity attributes recovered from the raw listings file (guests accommodated, bathrooms, beds); G4, three activity attributes from the same file (minimum nights, annual availability, and the host's total listing count); G5, two geospatial descriptors: the haversine distance to the city center, and the neighbourhood's listing count, a frequency encoding that replaces the high-cardinality neighbourhood attribute rather than one-hot expanding it, in line with evidence on encodings for high-cardinality categoricals [41]; and G6, the audited nightly price as a per-city log transform plus its standardized version (fit, like TF-IDF/SVD, once per city). The group labels G1–G6 are reused by the ablation of Section VII-F.

Across the three cities, materialization produced 33 matrix files in total: the per-city feature matrices above, the label vectors, and the intermediate text representations.

C. Model Families and Ladder Attempts

The ladder covers four GBDT families (HistGradientBoosting, XGBoost [18], LightGBM [19], and CatBoost [20]), evaluated first in the cold scenario (I01–I04), after which the strongest family (LightGBM) carries the feature-centric attempts: full (I05), text fusion (I06–I08), and engineered features (I13). Model-centric attempts then compose on the best preparation: out-of-fold stacking of four families with a logistic meta-learner (I11, I19) in the spirit of AutoGluon [29]; probability calibration with per-class decision thresholds tuned on an internal split (I12, I20); a zero-hurdle design chaining a zero-versus-rest classifier with the three-level model (I17) [36]; and TabPFN as a tuning-free foundation-model contender on GPU with context subsampling to 10,000 rows (I18) [15], [24]. Class weighting is a per-city policy decided on evidence, and SMOTE is deliberately excluded from the main ladder following [33], [34].

One evidence gate fired during execution: after dense E5 embeddings underperformed sparse TF-IDF in the cold scenario in all three cities (I08 versus I06; per-city deltas -0.51, -0.36, and -0.15 points, consistent in direction but comparable to per-fold dispersion, Section VII-C), the planned full-scenario E5 attempt (I09) was skipped as dominated; the decision is logged in the experimental record. The gate's economic logic is that the expected value of I09 was low (text of any kind adds at most +0.25 points in the full scenario, Section VII-C), while full+E5 is the most expensive preparation to search; we acknowledge that TF-IDF dominance in the full scenario is an extrapolation from the cold scenario, not a measurement. This is precisely the scenario against which Shi et al. [31] warn: n-gram baselines remain difficult to outperform.

D. Class Imbalance Policy

Tercile labeling yields near-balanced classes (Table II), but residual imbalance interacts with each market differently. Class weighting is therefore treated as a per-city, evidence-driven policy rather than a global switch: balanced weights added 0.5–1.3 points in the Buenos Aires and Santiago cold studies but were neutral to negative in Mexico City, and in the full-scenario Santiago studies the entire top-10 was unanimously unweighted. Synthetic oversampling is excluded from the main ladder on the strength of recent evidence that it does not help well-tuned strong learners and can distort calibration [33], [34]; the ladder instead evaluates the calibration-plus-thresholds route explicitly (I12, I20).

E. Hyperparameter Optimization

Each distributed attempt runs an Optuna [39] study per city under a fixed nominal trial budget (doubled in the contingency campaigns), combining a random exploration stage with TPE exploitation, coordinated through a central PostgreSQL storage and executed by 72 concurrent single-threaded workers, a massively parallel regime in the spirit of ASHA [40]. Search spaces follow the tunability literature [23], with targeted widenings applied only where earlier studies pressed top-10 configurations against a bound: the iteration cap was raised to 1,200 with a learning-rate floor of 0.003, XGBoost depth extended to 16, and CatBoost switched to lossguide growth for Mexico City only. Log-scaled parameters never saturated their ranges, so no blanket widening was applied.

F. Evaluation Protocol

Every configuration is scored by stratified five-fold cross-validation with macro F1 as the objective and macro precision and macro recall also recorded. Fold strata derive from global terciles; label thresholds from the training portion of each fold (Section IV). We report cross-validation point estimates with their fold dispersion, mindful that such intervals estimate the procedure's error rather than a single model's [38].

For the key attempt-level comparisons we additionally report two-sided *exact Wilcoxon signed-rank tests* over the fifteen paired fold scores (five folds $\times$ three cities; fold composition is identical across attempts by design), computed from the recorded per-fold scores, with Holm adjustment across the four comparisons reported in Table V [37]. Two caveats make these tests supporting rather than confirmatory evidence: they are post hoc (the confirmatory 10×3-fold protocol contemplated in the pre-specification was not executed), and fold scores within a city share training data, so the fifteen pairs are not fully independent. A paired comparison against the TabPFN attempt is omitted from Table V: its per-fold scores stem from a manual re-run outside the orchestrated pipeline (Section VII-E).

The same five-fold cross-validation serves both as the search objective that selects configurations and as the reported score, which introduces selection optimism into per-attempt nominal figures. Rather than nested cross-validation (infeasible at the scale of 27,219 registered trials), the study controls this with repeated outer resampling: four fresh fold shuffles re-evaluate the winning configurations held fixed, bounding the optimism at 0.2–0.3 points of macro F1 (Section VII-F).

For the best configurations, per-class precision, recall, and F1, confusion matrices, and order-sensitive metrics were obtained through a deterministic re-evaluation on identical folds (Section VII-F); the order-sensitive set comprises mean absolute error on class indices (low = 0, medium = 1, high = 2), quadratic weighted kappa, and the rate of two-level (low$\leftrightarrow$high) errors. However, these metrics did not guide any configuration selection.

VI. Experimental Setup

Experiments ran on a small heterogeneous cluster of commodity machines joined over a private overlay network: x86-64 Linux servers (one hosting the hyperparameter-search coordinator and one contributing a single consumer-grade GPU for TabPFN), together with Apple Silicon macOS nodes, one of which orchestrates the evidence ladder. Cross-architecture numeric discrepancy was measured at roughly 0.001 macro F1 or below (a paired re-execution reproduced a stacking attempt to within 0.0001; the revision-stage re-evaluation campaign observed a single fold at 0.0011), well under per-city fold dispersion (0.003–0.013).

All searches followed the same simple scheme: every attempt received an identical nominal trial budget, and the three contingency campaigns doubled it. A central registry recorded every trial; a few studies deviated from the nominal budget for operational reasons, and those deviations are part of the recorded experimental record. Only a fraction of the registered trials completed the full five-fold evaluation; the rest were pruned early, the expected behavior of a massively parallel search. Table IV reports the configuration and outcome of every attempt and contingency campaign.

Two operational observations shaped the program. Computational cost is dominated by the feature representation rather than the learner: text-based preparations are several times more expensive per trial, and the low-signal cold scenario pushes the search toward larger trees, raising its cost further. Across model families, CatBoost carried a substantial training premium over LightGBM while also ranking last among the four families, an operational argument, alongside accuracy, for adopting LightGBM as the default learner.

VII. Results

A. Overview of the Evidence Ladder

Table IV and Fig. 2 present the complete ladder, now with fold-level dispersion, and Table V the paired statistical comparisons. The nominally best pipeline is attempt I13 (LightGBM with HPO over the engineered *full_v2* preparation), with an average macro F1 of **0.8249** (Buenos Aires 0.8223, Santiago 0.8227, Mexico City 0.8299; averages here and throughout are computed on unrounded per-city values). Its +0.25-point margin over the previous leader I07 is smaller than the typical per-fold dispersion (0.003–0.013) but consistent in direction (12 of 15 paired folds; exact Wilcoxon $p = 0.013$, sitting exactly at the 0.05 boundary after Holm adjustment). We therefore describe I13 as the best attempt only *nominally*, within the fold-level variability of its closest competitors. The accurate summary is not that I13 separated from the field, but that no attempt did. The fifteen executed attempts plus three contingency campaigns converge to a plateau at 0.81–0.83, discussed below.

Fig. 3 details the per-city results: the cold-to-full gap dwarfs every within-full difference in all three markets, and the per-city bests within the pre-specified program are 0.8223 (Buenos Aires, I13), 0.8261 (Santiago, contingency C2), and 0.8299 (Mexico City, I13). The Santiago figure requires care with fold noise: C2's 0.8261 is statistically indistinguishable from I05's 0.8232 on the same folds (fold std 0.010–0.013; exact Wilcoxon on the five Santiago folds, $p = 0.19$), so we do not read it as evidence that doubling the budget helped.

B. The Cold-Start Gap as a Measure of Circularity

The distance between the best cold-start attempt (I03, average 0.5945) and the best full-scenario attempt (I13, 0.8249) is

TABLE IV. EVIDENCE LADDER: CONFIGURATION, AVERAGE MACRO F1 OVER THE THREE CITIES, AND FOLD DISPERSION $\bar{\sigma}$ (MEAN OVER CITIES OF THE PER-FOLD STD). ROWS FOLLOW EXECUTION ORDER. BOLD = BEST PRE-SPECIFIED THREE-CITY AVERAGE. C1–C3 REPORT THE *Target City's* MACRO F1, NOT A THREE-CITY AVERAGE (THEIR $\bar{\sigma}$ IS THAT CITY'S FOLD STD); ITALICS = SANTIAGO PER-CITY BEST WITHIN THE PRE-SPECIFIED PROGRAM. POST-HOC ANALYSES ARE REPORTED SEPARATELY IN TABLE VII (SECTION VII-F)

ID	Model / idea	Preparation	Macro F1	$\bar{\sigma}$
I01	HistGB	cold	0.5881	.0056
I02	XGBoost	cold	0.5934	.0056
I03	LightGBM	cold	0.5945	.0056
I04	CatBoost	cold	0.5848	.0063
I05	LightGBM	full	0.8199	.0075
I06	LightGBM	cold_tfidf	0.6136	.0063
I07	LightGBM	full_tfidf	0.8224	.0058
I08	LightGBM	cold_e5	0.6102	.0068
I09	(skipped by E5 gate)	full_e5	—	—
I11	Stacking OOF (4 fam.)	full_tfidf	0.8209	.0063
I12	Calibration + thresholds	full_tfidf	0.8145	.0072
I17	Zero-hurdle	full_tfidf	0.8190	.0050
I13	LightGBM + new features	full_v2	**0.8249**	.0047
I18	TabPFN (10k rows)*	full_v2	0.8205	.0062
I19	Stacking OOF (global)	full_v2	0.8238	.0054
I20	Final calibration	full_v2	0.8186	.0060
C1	Conting. BA (1,440 trials)	full_tfidf	0.8178	.0040
C2	Conting. CL (1,440 trials)	full	*0.8261*	.0101
C3	Conting. MX (1,440 trials)	full_tfidf	0.8271	.0047

*I18 failed in the orchestrator and was re-executed manually on the GPU node; its metrics stem from that manual run (Section VII-E).

TABLE V. EXACT WILCOXON SIGNED-RANK TESTS OVER THE 15 PAIRED FOLD SCORES (5 FOLDS × 3 CITIES); EXPLORATORY, POST HOC (SECTION V-F)

Comparison	ΔF1 (pts)	Folds won	p	p (Holm)
I07 vs. I05 (text on full)	+0.25	11/15	0.073	0.096
I13 vs. I07 (features vs. text)	+0.25	12/15	0.013	0.050
I13 vs. I05 (features on full)	+0.51	13/15	0.015	0.050
I06 vs. I08 (TF-IDF vs. E5, cold)	+0.34	11/15	0.048	0.096

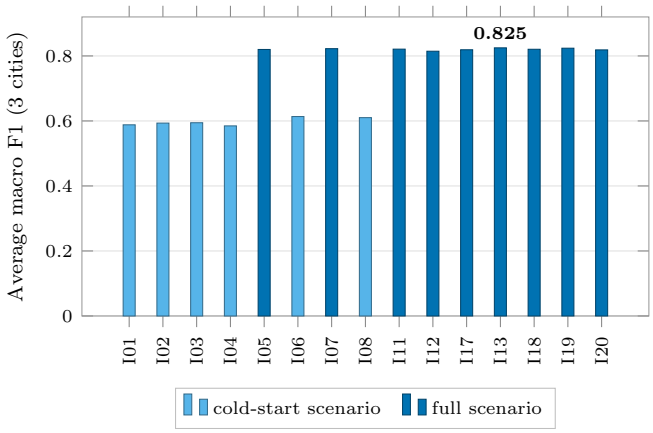

Fig. 2. Average macro F1 of each ladder attempt. Cold-start attempts saturate near 0.61; full-scenario attempts form a plateau at 0.81–0.83, with I13 (0.8249) the nominal best.

23.0 points of macro F1 (this is the definition used throughout; the distance from I03 to the center of the 0.81–0.83 plateau is about 22.5 points). The nominal cold set is, moreover, an optimistic upper bound: under the audited strict variant (P4/P5, Section VII-F) the best cold-start average drops to 0.5301 and

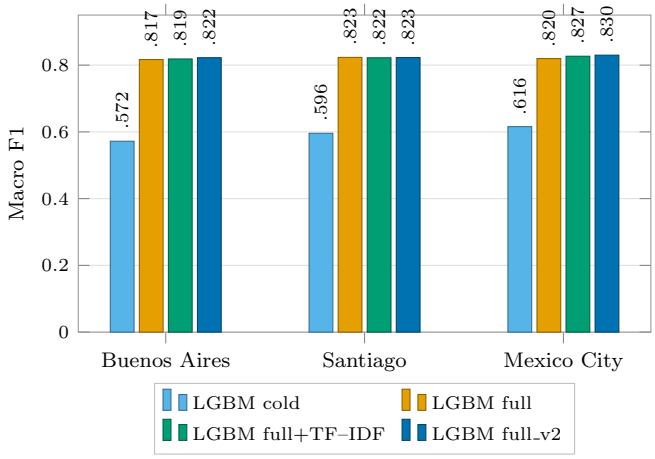

Fig. 3. Macro F1 per city for the key LightGBM configurations (cold I03, full I05, full+TF–IDF I07, full_v2 I13).

the gap widens to 29.5 points, so the circularity is larger, not smaller, than the pre-specified figure suggests.

This is the study's central substantive finding: review-derived features explain engagement retrospectively to a degree no creation-time feature set approaches, quantifying the circularity that is unavailable to any deployment-time predictor. Cold-start hyperparameter search is also saturated: all eighteen cold studies (six cold attempts × three cities) show top-1 to top-20 gaps below 0.003, and HPO adds only 1.0–1.6 points over defaults, so the cold-start frontier is unlikely to advance through additional search alone; new signal is required.

C. Marginal Value of the Spanish Description Text (RQ1)

Table VI isolates the text question, now with per-fold dispersion. Spanish descriptions vectorized with TF-IDF/SVD lift the cold scenario in all three cities: Buenos Aires 0.5720 to 0.5877 (+1.6 points), Santiago 0.5958 to 0.6198 (+2.4), and Mexico City 0.6157 to 0.6333 (+1.8). Each gain is a multiple of the corresponding fold std, empirical evidence that Spanish listing text carries engagement signal. In the full scenario the same fusion adds only +0.25 points on average (I07 = 0.8224 versus I05 = 0.8199; Mexico City +0.70, Buenos Aires +0.17, Santiago −0.12), a margin comparable to fold-level variability and not statistically distinguishable ($p = 0.073$ uncorrected, Table V), indicating that review-derived features already subsume most of the text's semantic content. Dense multilingual E5 embeddings *underperformed* sparse TF-IDF in all three cities in cold (I08 = 0.6102 versus I06 = 0.6136; per-city deltas −0.51, −0.36, −0.15 points). The embeddings demand roughly four times more computation and 384 rather than 100 dimensions, and the direction is consistent but the margin is within fold noise ($p = 0.048$ uncorrected, 0.096 after Holm). The operative conclusion is therefore cost-based: for short Spanish listing descriptions the low-cost n-gram baseline performs at least as well at a fraction of the computational cost, exactly the outcome benchmark studies caution about [31], and the evidence gate consequently skipped the full-scenario embedding attempt. The text block also places the best observed cold-plus-text performance at about 0.61: whatever

TABLE VI. PER-CITY MACRO F1 (MEAN AND PER-FOLD STD) OF THE TEXT-FUSION ATTEMPTS (LIGHTGBM)

City	Cold-start scenario			Full scenario		
	cold (I03)	+TF-IDF (I06)	+E5 (I08)	full (I05)	+TF-IDF (I07)	full_v2 (I13)
Buenos Aires	0.5720	0.5877	0.5826	0.8168	0.8185	0.8223
fold std	±.002	±.004	±.004	±.006	±.004	±.003
Santiago	0.5958	0.6198	0.6162	0.8232	0.8220	0.8227
fold std	±.010	±.010	±.009	±.013	±.009	±.008
Mexico City	0.6157	0.6333	0.6318	0.8197	0.8267	0.8299
fold std	±.004	±.005	±.007	±.004	±.004	±.003
Average	0.5945	0.6136	0.6102	0.8199	0.8224	0.8249

separates cold-start prediction from the 0.82 plateau is not recoverable from listing descriptions alone within the evaluated representations. An in-fold refit of the text transform (P6, Section VII-F) bounds the optimistic bias of the shared global fit at +0.16 to +0.66 points per city in this scenario (average +0.43), well below the text gains, so those gains are not artifacts of the global fit, although in-fold the sparse–dense margin itself becomes statistically indistinguishable.

D. Model-Centric Attempts and the Plateau (RQ2)

Every model-centric attempt scored at or below the tuned single LightGBM: four-family out-of-fold stacking 0.8209 (I11) and 0.8238 (I19); calibration with tuned per-class thresholds 0.8145 (I12) and 0.8186 (I20); the zero-hurdle chain 0.8190 (I17). Only the *feature-centric* attempt moved the frontier nominally: full_v2's engineered columns took 0.8224 to 0.8249 and produced the best per-city results in Buenos Aires and Mexico City. A leave-one-group-out ablation attributes that gain (Table IX, Section VII-F): every group contributes a sub-point share, with the activity block G4 the largest and most consistent. The result carries the statistical caveat of Section VII-A, and the top of the plateau is statistically flat (I13 versus I19: $p = 0.095$). Doubling the HPO budget did not improve results either: the doubled-budget Buenos Aires contingency reached 0.8178, below the city's standard-budget best, and the Santiago contingency's nominal best is within fold noise (Section VII-A). The convergence is broad. All 30 full-scenario per-city results land inside 0.8114–0.8299: the winners of the 12 full-scenario search studies (I05, I07, I13 and the three contingency campaigns) plus the 18 per-city evaluations of the remaining full-scenario attempts (I11, I12, I17, I18, I19, I20). We read this as an observed performance plateau within the evaluated feature and model space rather than as under-search; these experiments alone cannot attribute the plateau to the data, the feature representation, or the model class.

E. TabPFN Under the Pre-Specified Protocol (RQ3)

TabPFN with a 10,000-row context subsample and no tuning reached an average macro F1 of 0.8205 on a single consumer-grade GPU. This is within 0.5 points of the tuned LightGBM, above the calibration and hurdle attempts, and marginally below both stacking attempts (I11 0.8209, I19 0.8238), at a small fraction of the computational cost of the tuned pipelines. Its best city was Santiago (0.8260, essentially matching that city's best within the pre-specified program, 0.8261). One procedural caveat applies: I18 initially failed inside the orchestrator and was re-run manually on the GPU node; its metrics come from the manual run rather than the orchestrated per-trial pipeline. Under the pre-specified subsample, the conclusion was: competitive without tuning at a fraction of the cost, but not superior to a well-tuned GBDT.

F. Post-Hoc Analyses

We analyze P1/P2 shortly after its closure, and P3–P7, the ablation, and the resampling extension during revision. They are exploratory follow-ups outside the evidence ladder, and are reported apart from the pre-specified attempts (in Tables VII–IX rather than Table IV) to keep the two levels of evidence distinct. All follow-up analyses employ the fixed winning configurations and adhere to the same evaluation protocol, using identical folds except in cases where new resampling is explicitly required. When a follow-up analysis reassesses a previously registered attempt, the reproduced fold-level scores agree with the original registered values within a maximum deviation of 0.0011. The pre-specified conclusion regarding TabPFN remains unchanged from that established in Section VII-E.

1) Full-context TabPFN (P1) and multi-resampling confirmation (P2): A **post-hoc follow-up (P1)** removed the context subsample (all three cities fit within TabPFN-2.5's 50,000-row context), with the same protocol and data. It reached macro F1 0.8238±0.0037 (Buenos Aires), 0.8273±0.0103 (Santiago), and 0.8287±0.0026 (Mexico City), averaging **0.8266** (Table VII). This is the best three-city average observed in the study (the best pre-specified attempt, I13, reached 0.8249) and the best per-city result in Buenos Aires and Santiago. A **multi-resampling confirmation (P2)** then re-evaluated both sides under four fresh cross-validation resamplings (fresh fold shuffles, with every model configuration held fixed): full-context TabPFN averaged **0.8268±0.0004** against **0.8234±0.0005** for the program's best per-city LightGBM configurations, winning **all twelve city–resampling pairs** (mean margin $+0.0034$). The exercise also quantifies selection bias: the tuned configurations lose ≈0.3 points under resamplings other than the one on which they were selected (0.8261 → 0.8234), while the untuned foundation model is unaffected. At full context, TabPFN therefore *consistently* outperforms the program's best tuned GBDT configurations under fresh resampling; the earlier conclusion that it was not superior was an artifact of the context subsample. The finding remains post hoc, on a single corpus, and against fixed (not re-tuned per resampling) competitor configurations, consistent with the size regime reported by [15], [24].

2) Evaluation of selection optimism through repeated outer resampling: The four fresh resamplings of P2 were extended to the key tuned attempts, every configuration held fixed: I05 falls from 0.8199 (selection folds) to 0.8179 (fresh-resampling mean), I07 from 0.8224 to 0.8197, and I13 from 0.8249 to 0.8232 (between-resampling std 0.0004–0.0005). Selection optimism is therefore bounded at 0.2–0.3 points of macro F1 for every headline configuration, no corrected value leaves the 0.81–0.83 band, and the attempt ranking is unchanged: the plateau is not an artifact of selecting and reporting on the same folds.

TABLE VII. POST-HOC ANALYSES, OUTSIDE THE PRE-SPECIFIED PROTOCOL (SECTION VII-F): FULL-CONTEXT TABPFN PER CITY (P1) AND MULTI-RESAMPLING CONFIRMATION (P2)

P1: full-context TabPFN (5-fold CV)	Macro F1	fold std
Buenos Aires	0.8238	.0037
Santiago	0.8273	.0103
Mexico City	0.8287	.0026
Average	**0.8266**	
P2: four fresh resamplings (mean over cities)	**Macro F1**	**std**
TabPFN, full context, untuned	0.8268	.0004
Best tuned LightGBM configurations (fixed)	0.8234	.0005

P2: TabPFN wins 12 of 12 city–resampling pairs; mean margin $+0.0034$.

TABLE VIII. POST-HOC ORDINAL METRICS AND PER-CLASS F1, THREE-CITY AVERAGES (MAE: MEAN ABSOLUTE ERROR ON CLASS INDICES; QWK: QUADRATIC WEIGHTED KAPPA; e_2: RATE OF TWO-LEVEL ERRORS)

	Macro F1	MAE	QWK	e_2 (%)	F1 per class		
Configuration					low	medium	high
I03 (cold)	0.5945	0.460	0.581	6.16	0.690	0.445	0.649
I05 (full)	0.8199	0.177	0.869	0.06	0.909	0.721	0.829
I07 (full_tfidf)	0.8224	0.175	0.871	0.06	0.907	0.725	0.835
I13 (full_v2)	0.8249	0.173	0.873	0.05	0.907	0.729	0.838
P3 Frank–Hall	0.8232	0.174	0.872	0.04	0.907	0.727	0.836
P1 TabPFN full*	0.8263	0.170	0.875	0.04	0.908	0.730	0.841

*From the persisted re-evaluation; differs from Table VII by 0.0003 (TabPFN GPU inference is not bit-reproducible; fold-level deltas ≈0.002).

3) Ordinal evaluation and per-class error (P3): Table VIII reports order-sensitive metrics and per-class F1 for the key configurations, computed from persisted out-of-fold predictions on identical folds. Two facts stand out.

a) First, the flat multiclass models already respect the label order: in the full scenario, two-level (low↔high) errors affect only 0.04–0.06% of listings and quadratic weighted kappa is about 0.87, so misclassifications are overwhelmingly between adjacent levels. In the cold scenario the two-level error rate rises to 6.2%, quantifying in ordinal terms how much harder deployment-time prediction is.

b) Second, a genuine ordinal classifier does not improve on the flat approach: The Frank–Hall decomposition (P3), run with I13's LightGBM configuration on full_v2, averages 0.8232, inside the plateau and slightly below I13 (exact Wilcoxon over the fifteen paired folds, $p = 0.003$, I13 wins 12 of 15), with ordinal metrics indistinguishable from I13's.

c) The per-class breakdown answers which tier concentrates the residual error: the middle tercile, at F1 about 0.72–0.73 in the full scenario against 0.91 (low) and 0.84 (high), consistently across cities and model classes; in the cold scenario the middle tercile falls to 0.44. The confusion matrices confirm that the extreme classes, the actionable ones for platform operators, are recovered far more reliably than macro F1 alone suggests.

4) Feature-group ablation and attribution: A leave-one-group-out ablation on full_v2 re-evaluates I13's fixed configuration after removing each engineered feature group defined in Section V-B, using identical folds (Table IX). The observed gain is distributed across feature groups: each group contributes to performance, all six ablations produce drops below one macro-F1 point, and all remain statistically significant after Holm correction at the 0.05 level (with the weakest at $p = 0.041$). The G4 activity block (minimum nights, annual availability, and host listing count) emerges as the largest and most consistent contributor (-0.45 macro-F1 points, with a performance decrease in all 15 folds and a Holm-adjusted $p = 0.0004$); removing all 18 engineered features reduces the resulting macro-F1 score to 0.8196.

TABLE IX. POST-HOC LEAVE-ONE-GROUP-OUT ABLATION OF FULL_V2'S ENGINEERED GROUPS (I13 CONFIGURATION FIXED; THREE-CITY AVERAGE; EXACT WILCOXON OVER 15 PAIRED FOLDS, HOLM-ADJUSTED)

Removed group	Macro F1	Δ (pts)	Folds lost	p (Holm)
G1 amenities (6)	0.8227	-0.22	10/15	0.037
G2 word counts (2)	0.8223	-0.26	12/15	0.006
G3 capacity (3)	0.8231	-0.18	13/15	0.037
G4 activity (3)	0.8204	-0.45	15/15	0.0004
G5 geospatial (2)	0.8229	-0.20	11/15	0.041
G6 price (2)	0.8226	-0.23	11/15	0.021
All 18 (control)	0.8196	-0.53	—	—

The feature-attribution analysis yields a consistent pattern: LightGBM gain importances, recomputed independently for each fold, rank the review-derived features highest, followed by instant bookability, 90-day availability, and amenity count, while minimum nights and host listing count, both from G4, rank as the most influential engineered features.

Feature importances remain stable across folds (mean Spearman correlation = 0.80; top-15 Jaccard similarity = 0.86) and show moderate consistency across cities (mean Spearman correlation = 0.53), lending support to this interpretation despite the relatively small effect sizes. All ablated variants retain I13's original hyperparameters, a conservative design choice that may slightly overestimate the performance loss associated with removing each feature group.

5) Strict cold-start after the temporal audit (P4, P5): Re-running the cold-start attempts without the three T3 features of Table III (frozen configurations, identical folds) lowers I03 from 0.5945 to 0.5301 (P4; per city 0.5175, 0.5324, 0.5405) and I06 from 0.6136 to 0.5762 (P5). The deployable cold-start reference is therefore roughly 0.53–0.58, and the circularity gap of Section VII-B widens to 29.5 points. Notably, the marginal value of the Spanish description grows once the dynamic features are removed: text adds +4.0, +4.5, and +5.3 points per city over the strict structured set, versus +1.6 to +2.4 nominally. Where dynamic snapshot signals are unavailable, the listing text carries correspondingly more of the recoverable signal. A moderate verification search over the original hyperparameter space (TPE, seed 42; 50 and 35 fresh trials per city on the strict preparations) confirms the frozen-configuration figures: its best scores differ from them by -0.2 to $+0.5$ points per city, so the strict drop is not an artifact of inherited hyperparameters and the widened gap stands.

6) Quantifying the global-fit bias (P6): Refitting the TF-IDF/SVD transform (and full_v2's standardized log price) inside each training fold, with winning configurations frozen, bounds the optimistic bias of the shared global fit: the globally fitted scores exceed their fold-wise refitted counterparts by 0.43 points of macro F1 on average for I06 (cold_tfidf; per

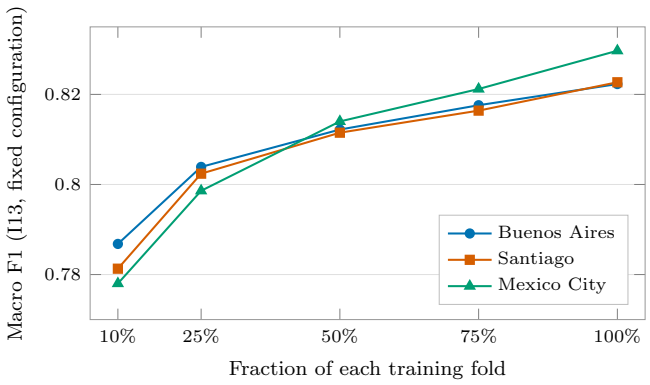

Fig. 4. Post-hoc learning curves (P7): Macro F1 of I13's fixed configuration trained on stratified fractions of each training fold (identical folds, seed 42). The curves decelerate but do not fully flatten.

city 0.16, 0.66, 0.47), by 0.29 for I07, and by 0.10 for I13, the last within fold noise (exact Wilcoxon $p = 0.30$). The E5 attempt I08, whose embeddings require no corpus-level fitting, reproduces its registered scores exactly (fold-level delta = 0.000), providing an internal validation of the refitting procedure. The estimated bias is therefore several times smaller than the performance gains attributed to text features in the cold scenario and negligible in the full scenario: none of the internal model comparisons or conclusions regarding the performance plateau are altered, and the absolute cold-scenario text scores exhibit at most 0.66 macro-F1 points of optimistic bias in the most affected city–attempt pair (0.43 on three-city average). The refit does temper one comparison: under fold-wise refitting, sparse TF-IDF and dense E5 no longer exhibit a statistically significant difference (three-city averages 0.6093 versus 0.6102; exact Wilcoxon over the fifteen paired folds, $p = 0.49$), so the sparse-versus-dense preference rests on cost, not accuracy.

7) Learning curves (P7): Fig. 4 evaluates I13's fixed configuration across stratified fractions of each training fold. The curves exhibit a marked deceleration, with progressively smaller gains as the training-set size increases, but have not yet fully flattened at the current sample sizes. The final quarter of the available training data still yields gains of approximately 0.5–0.8 macro-F1 points, depending on the city, and the observed trend suggests that a further doubling of the training data would yield an improvement of at most approximately one macro-F1 point. Sample size therefore remains a relevant dimension of the evaluated space: the plateau identified in Section VII-D should be interpreted as specific to the current corpus and sample size, and additional listings, similarly to the incorporation of new feature sources, may still produce measurable performance gains.

G. Operational Findings

CatBoost carried a substantial per-study training-cost premium over LightGBM and ranked last among the four families in cold (0.5848), relegating it to a diversity source for stacking. Santiago shows 2–3 times the fold variance of the other cities (standard deviations 0.008–0.013; fold range 0.807–0.841 in I05), so sub-point deltas there are not meaningful and all Santiago comparisons require per-fold pairing. Mexico City is the most expensive market to search when text is present, while the compact *full* preparation is by far the cheapest to search.

VIII. Discussion

1) Interpretation of the plateau: The convergence of all 30 full-scenario per-city results (from heterogeneous families, preparations, ensembles, and budgets) into a two-point band around macro F1 0.82 constitutes an observed performance plateau within the evaluated feature and model space, stronger evidence than any single result.

These experiments cannot identify where the limit lies: in the information the data carry, in the feature representation, or in the model classes evaluated.

The ladder itself probes the representation axis as a feature-addition sequence (full 0.8199, full_tfidf 0.8224, full_v2 0.8249), where each added block contributes at most +0.25 points, a pattern of diminishing returns that maps the evaluated space without exhausting it. Within the pre-specified program no learning curves or label-noise bounds were computed, and the convergence of correlated tree ensembles on near-identical features could reflect a limit of this model class and feature resolution as much as one of the data. That TabPFN, a different model class, lands in the same band (indeed at its top in the post-hoc full-context follow-up, 0.8266; Section VII-F) narrows, though does not settle, the question.

If the limit does lie in the data, the residual error is plausibly dominated by information the platform snapshot does not contain (pricing dynamics, availability windows, external seasonality, off-platform demand shocks), an explicitly untested hypothesis. For practitioners the reframing stands either way: the next point of macro F1 is more likely to come from new data sources or richer representations than from additional compute over the current ones.

The post-hoc learning curves sharpen the reading: they decelerate strongly yet still rise at full data (P7, Section VII-F), so sample size remains an open axis alongside representation and model class.

The plateau also serves as a calibration reference for the field, since no comparable published figures exist for this formulation (Section II).

2) Feature engineering over model-centric refinement: The only intervention that moved the frontier, nominally and at the boundary of statistical distinguishability, was feature engineering, consistent with recurring findings in the tabular literature [21]. Stacking, calibration, and hurdle designs (each with prior evidence of gains in other regimes [29], [36]) were neutral or negative here, plausibly because a tuned LightGBM on a plateau leaves them little room for improvement, and calibration's internal splits shrink effective training data.

The post-hoc ablation adds attribution: the engineered-feature gain is diffuse across all six groups, led by the activity block, with importances stable across folds: evidence that the feature-centric margin, though at the boundary of statistical distinguishability, is systematic rather than noise (Section VII-F).

3) Sparse lexical representations over dense embeddings: The text result is a practical guideline for Spanish-language P2P corpora: short, formulaic listing descriptions are captured adequately by lexical overlap, at a fraction of the computational and dimensional cost of dense semantic representations, and the embeddings showed no compensating accuracy gain; under the in-fold refit the two representations are statistically indistinguishable (Section VII-F), so the preference for TF-IDF rests on cost, not accuracy. Embedding pipelines should therefore be justified against this baseline before deployment.

4) Cost profile of the tabular foundation model: A tuning-free model within half a point of the winning attempt (obtained on a single consumer-grade GPU at a small fraction of the computational cost of the distributed search program) makes foundation models a natural first baseline for future tabular studies in this domain, while the observed plateau itself remains unchanged.

5) Comparison with the binary antecedent: Kirkos reports F-measure 0.792–0.835 for a *binary* mean-split task in a single European city [13]. The classes, metric variant, market, features, and era all differ, so the figures are not comparable and we cite them as context only; indeed, no directly comparable published figure exists (Section II), which is precisely the gap this study addresses. Our three-level task is structurally harder (macro averaging penalizes every class equally, and the middle tercile borders both neighbors), and the per-class analysis of Section VII-F confirms that the middle tercile concentrates the residual error (full-scenario F1 about 0.73, versus 0.91 and 0.84 for the extremes), with confusions overwhelmingly between adjacent levels.

6) Practical implications: For platforms and hosts, the actionable reference is the observed cold-start range: at listing creation, three-level engagement is predictable at macro F1 of roughly 0.59–0.63 within the evaluated space, and at 0.53–0.58 under the audited strict set, since snapshot-measured availability, superhost status, and tenure inflate the nominal figure (Section VII-F). Adding a well-written Spanish description helps more the stricter the setting: up to +2.4 points nominally, and +4.0 to +5.3 points over the strict structured set, via low-cost lexical features. Per-class results add operational nuance: the extreme classes, the ones that drive ranking, trust, and regulatory attention, are recovered at F1 about 0.91 (low) and 0.84 (high), while the middle tercile absorbs most confusion (0.73), so decisions keyed to the extremes are more reliable than the macro average suggests. For market observatories and regulators working retrospectively, the full-scenario models classify at 0.81–0.83 with per-city calibration choices (class weights help in Buenos Aires and Santiago, not in Mexico City). And for research budgeting, the ladder shows that a study of this scope is feasible on a small cluster of commodity machines, with the single largest saving coming from an evidence gate that automatically skipped a dominated attempt.

IX. Conclusion and Future Work

This study delivered the first three-level engagement classification of peer-to-peer home-sharing listings in Latin America, on 76,624 listings across Buenos Aires, Santiago, and Mexico City. It mapped the attainable performance landscape with a pre-specified, amendment-logged, evidence-gated ladder of 27,219 registered trials in 30 hyperparameter studies. Four findings summarize the program. A tuned LightGBM over engineered features achieves macro F1 0.8249 on average (up to 0.8299 in Mexico City). The cold-start gap of 23.0 points quantifies the circularity of review-derived features; the audited strict variant widens the deployable gap to 29.5 points. Spanish text helps where information is scarce, increasingly so the stricter the setting, with sparse TF-IDF matching multilingual embeddings at a fraction of the cost. An untuned tabular foundation model matches the tuned pipelines with no hyperparameter search at all; a post-hoc, exploratory follow-up at full context (P1), upheld under four fresh resamplings (P2), suggests the untuned model can exceed the best pre-specified average (0.8266 versus 0.8249; twelve of twelve paired comparisons, mean margin $+0.0034$), though this comparison lies outside the pre-specified protocol (Section VII-F). The post-hoc analyses complete the picture: ordinal metrics and a Frank–Hall baseline show that the flat models already respect the label order, the middle tercile concentrates the residual error, and repeated outer resampling bounds selection optimism at 0.2–0.3 points. The stable 0.81–0.83 band across all full-scenario per-city results constitutes an observed performance plateau, confined to the feature and model space evaluated here; whether it reflects a limit of the data, of the feature representation, or of the model families evaluated remains an open question.

Several limitations bound these findings. Review counts proxy engagement rather than observed bookings [2], and the corpus is a cross-sectional snapshot of three cities, scraped in December 2024 (Santiago), January–February 2025 (Buenos Aires), and June–July 2025 (Mexico City), so temporal generalization and other Latin American markets remain untested; listing descriptions are the only text exploited. The ordinal evaluation and the Frank–Hall baseline are post-hoc additions made in revision, limited to a single ordinal method with inherited hyperparameters, and macro F1 remains the pre-specified primary metric. The TF-IDF/SVD transform and the per-city log-price standardization were fit once per city over all rows; the post-hoc in-fold refit bounds the resulting optimism at +0.1 to +0.4 points on average (up to +0.7 in a single city) and leaves internal comparisons unaffected, though it also renders the sparse-versus-dense text comparison statistically indistinguishable. Inside Airbnb preserves no listing-creation timestamp, so the temporal audit classifies features by documented mutability, and the retained quasi-static features could still have been edited after creation. The same cross-validation selects configurations and reports their performance; nested cross-validation was not run, and the repeated-outer-resampling bound of 0.2–0.3 points covers configuration selection, not the full search process. Per-class breakdowns, ordinal metrics, and ablations come from revision-stage re-evaluations of fixed configurations, with ablated variants not re-tuned (a conservative bias). The statistical tests remain exploratory and post hoc, the full-context TabPFN follow-up and its confirmation remain comparisons on a single corpus against fixed competitor configurations, and the evidence gate that skipped I09 extrapolated TF-IDF's cold-scenario standing to the full scenario without measuring it. Finally, the protocol was amendment-logged rather than externally pre-registered, and Santiago's high fold variance widens all its comparisons [38].

Future work follows directly: richer ordinal methods beyond Frank–Hall (CORAL/CORN-style learners, learned thresholds), temporal signal from calendar and availability data, spatial modeling over neighborhood geometries, pooled multi-city training with market indicators to tame fold variance, guest-review text and image multimodality, conformal prediction over the ordered levels, larger corpora to extend the learning curves, a longitudinal multi-snapshot validation of the temporal audit, and a fully symmetric TabPFN–GBDT comparison in which the trees are re-tuned under each resampling.

REFERENCES

[1] Inside Airbnb, "Inside Airbnb: Get the data," 2026, https://insideairbnb.com/get-the-data, accessed July 2026.

[2] E. Ert, A. Fleischer, and N. Magen, "Trust and reputation in the sharing economy: The role of personal photos in Airbnb," *Tourism Management*, vol. 55, pp. 62–73, Aug. 2016.

[3] K. Xie and Z. Mao, "The impacts of quality and quantity attributes of Airbnb hosts on listing performance," *International Journal of Contemporary Hospitality Management*, vol. 29, no. 9, pp. 2240–2260, Sep. 2017.

[4] H. Zhang, F. J. Zach, and Z. Xiang, "Optimal distinctiveness of short-term rental property design," *International Journal of Hospitality Management*, vol. 120, p. 103737, 2024.

[5] B. Biswas, P. Sengupta, and D. Chatterjee, "Examining the determinants of the count of customer reviews in peer-to-peer home-sharing platforms using clustering and count regression techniques," *Decision Support Systems*, vol. 135, p. 113324, Aug. 2020.

[6] P. Sengupta, B. Biswas, A. Kumar, R. Shankar, and S. Gupta, "Examining the predictors of successful Airbnb bookings with Hurdle models: Evidence from Europe, Australia, USA and Asia-Pacific cities," *Journal of Business Research*, vol. 137, pp. 538–554, 2021.

[7] D. Chen, W. Zhang, J.-W. Bi, H. Qiu, and J. Lyu, "Hosts' online affinities and their impacts on the number of online reviews on peer-to-peer platforms," *Tourism Management*, vol. 100, p. 104817, Feb. 2024.

[8] L. Zhan and M. Cheng, "The projected identity of Airbnb hosts: A multimodal approach," *Journal of Hospitality and Tourism Management*, vol. 65, p. 101328, Dec. 2025.

[9] P. Rezazadeh Kalehbasti, L. Nikolenko, and H. Rezaei, "Airbnb price prediction using machine learning and sentiment analysis," 2019, arXiv:1907.12665.

[10] O. R. Priambodo and A. Sihabuddin, "Extreme learning machine with silhouette index for Airbnb price prediction," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020.

[11] D. Wang and J. L. Nicolau, "Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com," *International Journal of Hospitality Management*, vol. 62, pp. 120–131, Apr. 2017.

[12] K. Gyódi and Ł. Nawaro, "Determinants of Airbnb prices in European cities: A spatial econometrics approach," *Tourism Management*, vol. 86, p. 104319, 2021.

[13] E. Kirkos, "Airbnb listings' performance: Determinants and predictive models," *European Journal of Tourism Research*, vol. 30, p. 2142, 2022.

[14] A. Subroto and M. Christianis, "Rating prediction of peer-to-peer accommodation through attributes and topics from customer review," *Journal of Big Data*, vol. 8, no. 1, p. 9, Dec. 2021.

[15] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, pp. 319–326, 2025.

[16] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*, 2nd ed., ser. Econometric Society monographs. Cambridge: Cambridge University Press, 2013, no. 53.

[17] J. M. Hilbe, *Negative binomial regression*, 2nd ed. Cambridge: Cambridge University Press, 2012.

[18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[19] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 3146–3154.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 6638–6648.

[21] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022.

[22] D. McElfresh, S. Khandagale, J. Valverde, V. Prasad C., G. Ramakrishnan, M. Goldblum, and C. White, "When do neural nets outperform boosted trees on tabular data?" in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

[23] P. Probst, A.-L. Boulesteix, and B. Bischl, "Tunability: Importance of hyperparameters of machine learning algorithms," *Journal of Machine Learning Research*, vol. 20, no. 53, pp. 1–32, 2019.

[24] Prior Labs, "TabPFN-2.5: Advancing the state of the art in tabular foundation models," 2025, arXiv:2511.08667.

[25] Q. Li, R. Dong, N. Miklaucic, J. Hu, S. S. Omee, L. Wei, S. Dey, M. Hu, and J. Hu, "In context learning foundation models for materials property prediction with small datasets," *npj Computational Materials*, 2026.

[26] S. Ruiz-Villafranca, J. Roldán-Gómez, J. M. Castelo Gómez, J. Carrillo-Mondéjar, and J. L. Martínez, "A TabPFN-based intrusion detection system for the industrial internet of things," *The Journal of Supercomputing*, vol. 80, pp. 20 080–20 117, 2024.

[27] D. Zhao, Y. Zhu, Z. Wu, Y. Wan, X. Liu, S. Li, H. Xu, T. Hou, and C.-Y. Hsieh, "Revisiting ADMET prediction reliability under real-world challenges in the foundation model era," *Journal of Cheminformatics*, vol. 18, no. 1, p. 95, 2026.

[28] Z. Xu, C. Li, C. Guan, L. Xu, X. Wang, S. Jiang, N. Zhang, M. Gu, Y. Xin, and Y. Xu, "Development and external validation of an interpretable early prediction model for acute kidney injury using TabPFN and routine admission data: a retrospective cohort study," *BMC Medical Informatics and Decision Making*, vol. 26, no. 1, p. 242, 2026.

[29] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola, "AutoGluon-Tabular: Robust and accurate AutoML for structured data," 2020, arXiv:2003.06505.

[30] S. Jain, D. Proserpio, G. Quattrone, and D. Quercia, "Nowcasting gentrification using Airbnb data," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–21, 2021.

[31] X. Shi, J. Mueller, N. Erickson, M. Li, and A. J. Smola, "Benchmarking multimodal AutoML for tabular data with text fields," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

[32] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 text embeddings: A technical report," 2024, arXiv:2402.05672.

[33] Y. Elor and H. Averbuch-Elor, "To SMOTE, or not to SMOTE?" 2022, arXiv:2201.08528.

[34] R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster, "The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression," *Journal of the American Medical Informatics Association*, vol. 29, no. 9, pp. 1525–1534, 2022.

[35] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Proceedings of the 12th European Conference on Machine Learning (ECML)*, 2001, pp. 145–156.

[36] J. M. Rožanec, B. Fortuna, and D. Mladenić, "Predicting zero-inflated demand with a two-fold machine learning approach," 2023, arXiv:2310.08088.

[37] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[38] S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: What does it estimate and how well does it do it?" *Journal of the American Statistical Association*, vol. 119, no. 546, pp. 1434–1445, 2024.

[39] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, pp. 2623–2631.

[40] L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, J. Ben-Tzur, M. Hardt, B. Recht, and A. Talwalkar, "A system for massively parallel hyperparameter tuning," in *Proceedings of Machine Learning and Systems (MLSys)*, 2020.

[41] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," *Computational Statistics*, vol. 37, pp. 2671–2692, 2022.