Can Yelp become Restaurant Health Inspection Predictor?

Pre-Inspection Review Text and New York City Restaurant Health Scores

A Time-Ordered Study of 7,480 Yelp Reviews and 465 Matched Listings

Bob Zhao
The Webb Schools

Abstract

Where yelp reviewers see dinner; health inspectors see violations. This paper asks whether the first group can help predict the second without accidentally using information that arrived after inspection day. I reconstructed dated Yelp review text and review-level stars from a CC BY archive matched to New York City inspection scores. After removing 11 exact duplicate rows and requiring at least three earlier reviews, the primary sample contains 465 listings, 7,480 pre-inspection reviews, and 59 inspections with scores of 14 or more—the official boundary outside the 0–13 A-point range. Five models were compared in five repeats of five-fold cross-validation grouped by normalized address: a prevalence baseline, review stars plus volume, a matched four-variable review-history baseline, that same metadata plus a prespecified sanitation lexicon, and the same metadata plus 3,000 TF-IDF features. The metadata-plus-text model's AUROC was 0.503, versus 0.488 for its matched non-text baseline (paired difference 0.015; conditional 95% interval -0.016–0.043). The leaner stars-plus-volume model reached 0.518. Text average precision was 0.134 against a 12.7% event rate, and its Brier score (0.113) was slightly worse than the prevalence-only baseline (0.111). A crawl-time 64-feature reconstruction reached AUROC 0.711 when fit and scored on the same rows, then fell to 0.480 on unseen address groups. The open archive therefore supports an instructive negative result: these reviews contain interesting associations, but not stable evidence of incremental inspection-triage value under this stricter test.

Keywords: online reviews; restaurant inspections; text as data; temporal leakage; public health; model validation

1 Background

1.1 Two kinds of expertise walk into a restaurant

A diner and a health inspector can visit the same restaurant and observe different worlds. The diner notices taste, service, price, atmosphere, and occasionally something alarming enough to type in all caps. The inspector uses a formal checklist, looks behind equipment, checks temperatures, and assigns violation points. Neither view is complete. The computational-sociology question is whether many informal observations can become a useful public signal when they are aggregated, even though the observers were never trained or randomly assigned.

New York City's system makes the prediction target unusually concrete. Inspection violations carry points, and lower totals are better. Scores from 0 through 13 fall in the A range, 14 through 27 in the B range, and 28 or more in the C range [1]. The archive used here does not identify the inspection stage or prove which placard a customer saw, so I predict the score boundary rather than claiming that every record represents a posted B. That small wording choice matters: public data are not improved by giving them a more dramatic name.

1.2 Crowds are sensors, but very opinionated sensors

Online reviews have attractive properties for public-health surveillance. They are frequent, cheap to collect, and written between official visits. They may mention pests, dirty bathrooms, vomiting, or food left cold. A regulator with limited inspectors might use such language to rank establishments for earlier attention. Sadilek and colleagues showed that social-media signals can support a prospective inspection deployment, although their system used Twitter in Las Vegas rather than Yelp in New York [5]. This is a useful precedent, not a permission slip to assume every platform-city combination works.

The weaknesses are equally important. Yelp users choose where to eat, whether to review, what to mention, and how strongly to say it. Restaurants also differ in price, cuisine, neighborhood, and platform visibility. Platform filtering and competitive incentives further complicate the record; Yelp-filtered reviews are unusually extreme and are associated with business incentives, although a filtered review is not proof of fraud [6]. Reviews are therefore social behavior as much as measurement. Treating them like thermometer readings would be convenient. It would also be wrong.

1.3 What earlier studies found

Schomberg and colleagues built a keyword-and-tag model from Yelp pages and reported that it distinguished New York restaurants above their chosen inspection threshold with an AUROC of 0.77 [2]. Their public archive contains the raw material for this study [3]. The original project was framed largely as nowcasting: reviews near the present might supplement occasional inspections. Its released file, however, mixes inspection dates with crawl-time review aggregates, and it does not include the final coefficient vector needed for an exact independent replay. That creates a narrower question the archive can answer: what survives if every review feature is rebuilt using only text dated before the corresponding inspection?

A newer New York study by Farronato and Zervas used a much larger proprietary Yelp record and detailed inspection violations. It found that reviews were more informative about hygiene dimensions consumers can directly experience, such as pests and food temperature, than about less visible violations [4]. That work also studied demand and restaurant responses. The present paper does not attempt to out-scale it. Instead, it asks a reproducibility question that can be answered with an openly deposited file and methods simple enough to audit line by line.

1.4 Research questions

The study has three questions:

1. Do strictly pre-inspection stars, review volume, sanitation language, or general review text predict whether the next archived score lies outside the A-point range?
2. How different does performance look when a broad crawl-time feature set is scored on the rows used to fit it versus genuinely unseen address groups?
3. Among restaurants already in the A range, can review text recover meaningful variation in the 0–13 point score that the letter grade compresses?

The first question is the main test. The second separates a time problem from an evaluation problem. The third asks whether a coarse public label hides a continuous signal. None is causal. A Yelp complaint does not assign violations, and an inspection does not randomly assign complaints.

1.5 A design I deliberately did not use

A regression-discontinuity design around 13 versus 14 points sounds clever because the grade rule changes at that boundary. It is not defensible here. The archive has 57 observations at score 13 and only five at 14, lacks inspection type and placard-posting dates, and mixes A, B, C, and ungraded records. The score can also trigger reinspection rather than instantly reveal a stable posted grade. A method does not become causal merely because a cutoff is nearby. The threshold is used only to define a policy-relevant outcome.

2 Data

2.1 The public archive

The data are the New York City Dataset deposited on Figshare by Schomberg, Haimson, Hayes, and Anton-Culver [3]. Figshare labels the deposit CC BY 4.0. The CSV contains 743 matched Yelp-listing and inspection rows. For each row, a raw field stores HTML fragments for the listing address, dated Yelp review snippets, star images, and review text. Separate columns contain an inspection date, score, archived grade, ZIP code, crawl-time platform summaries, and keyword counts. Inspection dates run from March 25, 2014, through May 20, 2015.

Figure 1: Construction of the analytical sample. Review text is retained only when its publication date is strictly earlier than the matched inspection date.

Table 1: Sample construction and temporal audit

Stage	Retained	Excluded or comparison
Rows in deposited CSV	743	—
After exact-duplicate removal	732	11
Rows with at least one earlier review	597	135
Primary rows with at least three	465	132
Primary rows with score ≥14	59	406 score 0–13
Reviews in primary pre-period	7,480	1,148 later + 10 same-day

Notes: The source contains 9,076 review texts before exact-row deduplication. Eleven reviews dated on inspection day are excluded from the pre-period because their ordering within the day is unknown.

2.2 Reconstructing the reviews

The raw HTML is not a ready-made review table, so I parsed it rather than trusting the archive-level aggregates. For each listing, date tags and review paragraphs were extracted in page order. The first star image is the crawl-time listing average; the next N star images align with the N dated reviews; later star images belong to page widgets and were discarded. Date and text counts match in 731 of 732 deduplicated rows. The single mismatch was truncated to the shorter aligned length. Every row had enough star images for its aligned reviews.

A review entered the main predictors only if its date was strictly earlier than the inspection date. This removed 1,280 post-inspection texts from the deduplicated archive; 286 rows contained at least one. Within the primary 465 listings, 7,480 reviews are earlier and 1,148 are later. Stars and volume were recomputed from the review-level records. The archive's crawl-time Star, RC, and keyword totals were reserved for one explicitly labeled diagnostic. This prevents a quiet switch from forecasting to hindsight.

2.3 Outcome: outside the A-point range

$$Y_i = 1(score_i \geq 14)$$

The primary event is an inspection score at or above 14. It occurs in 59 of 465 listings, or 12.7%. I call this outcome outside the A-point range. I do not call it food poisoning, an unsafe restaurant, or a posted B, because the file cannot support those stronger statements. As a sensitivity check, I repeat the model with the earlier paper's strict score-greater-than-14 rule.

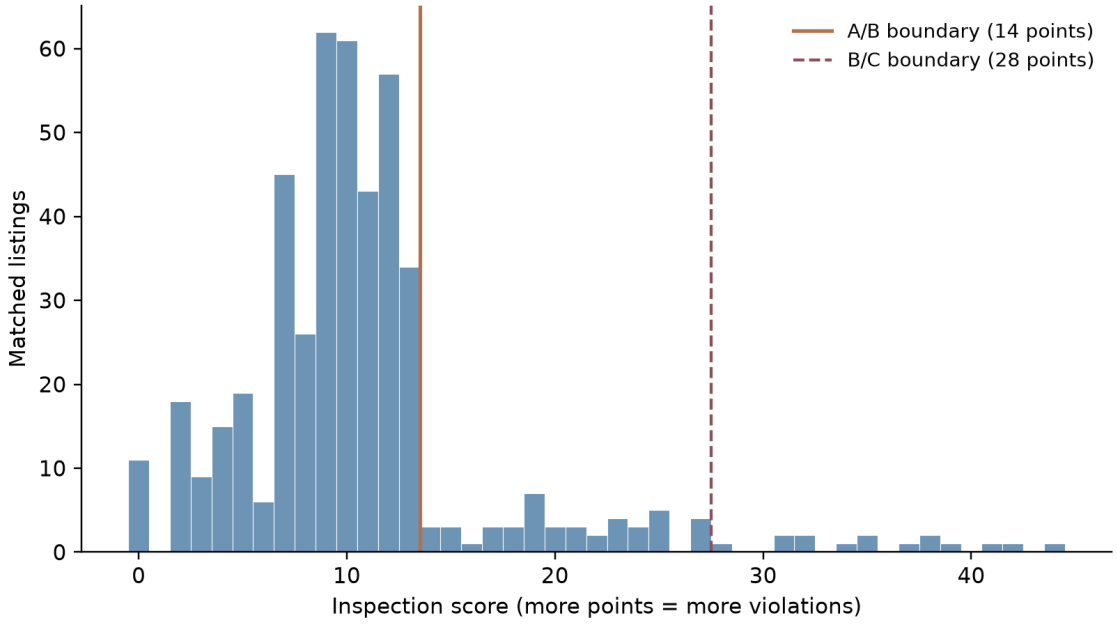

Figure 2: Inspection-score distribution in the primary sample. New York assigns more points for more violations; the solid line marks the first score outside the 0–13 A range.

2.4 What is actually sampled

The unit is an archived Yelp listing matched by the original researchers to an inspection record, not a random New York restaurant. Exact duplicates were removed, and six remaining pairs sharing a normalized address were forced into the same validation fold. The Yelp pages contain at most 40 selected reviews. The primary listing has a median of 14 earlier reviews, a median mean rating of 3.44 stars, and a median review span of 1,596 days. A target-blind sensitivity filter requiring at least half of a listing's reviews to contain food or dining vocabulary retains 440 listings and 57 of the 59 events.

3 Methods

3.1 Five models and one nosy look-ahead

The model ladder is designed to answer an incremental question. A complicated text model is useful only if it improves on information Yelp already displays in simpler form. Each logistic model estimates the probability that Y equals one. All learned transformations are fit inside the training fold.

Table 2: Prediction models

Model	Inputs	Uses later text?	Purpose
Prevalence only	Training-fold event rate	No	Sanity baseline
Stars + volume	Mean earlier stars; log earlier count	No	Simple Yelp signal
Review-history metadata	Stars, count, mean length, time span	No	Matched non-text baseline
Metadata + lexicon	Same metadata + 10 theme features	No	Auditable language test
Metadata + TF-IDF	Same metadata + up to 3,000 phrases	No	Flexible language test
All-review TF-IDF	Same design using all archived texts	Yes	Hindsight diagnostic only

Notes: All learned models use ridge-penalized logistic regression (L2 penalty, C = 1); the prevalence row is a constant baseline. The all-review model is never presented as a deployable forecast.

3.2 The transparent dictionary

Five word families were specified before inspecting model results: visible-hygiene complaints; illness language; cleanliness praise; negative experience; and positive experience. Each family produces two features: the log density per 1,000 words and the fraction of a listing's reviews containing at least one term. Review fractions keep a single long essay from automatically looking like a crisis. The sanitation families contain words such as roach, mice, dirty, bathroom, sick, vomit, diarrhea, and clean. The general experience families act as comparison signals rather than inspection facts.

3.3 The flexible text model

The TF-IDF model lowercases text, removes accent variation, uses one- and two-word phrases appearing in at least three training listings, and keeps at most 3,000 features. Sublinear term frequency prevents a word repeated ten times from counting ten times as loudly. Numeric predictors are standardized using training data. Ridge regularization shrinks correlated coefficients rather than selecting a fragile handful of magic words.

$$logit[P(Y_i = 1)] = \alpha + x_i'\beta + t_i'\gamma$$

Here x contains mean stars, log review count, mean review length, and review-history span; t contains either dictionary features or TF-IDF values. The review-history model uses x alone, making it the matched baseline for asking what the words add. The coefficients are prediction devices, not causal effects. If sauce receives a positive coefficient, the equation has not discovered a sauce violation.

3.4 Validation: the model must meet a restaurant it has not seen

The primary evaluation uses five repeats of five-fold stratified cross-validation. Normalized address is the grouping variable, so two listings at the same address never land on opposite sides of a fold. Every listing receives one held-out probability in each repeat, and the five probabilities are averaged. Vocabulary, inverse-document frequencies, missing-value medians, scaling, and coefficients are estimated using training rows only. This is the difference between testing memory and testing transport.

3.5 Metrics that do not reward doing nothing

Because only 12.7% of primary rows lie outside the A range, ordinary accuracy is misleading. A rule that predicts A-range for everyone is 87.3% accurate and finds zero high-score inspections. I therefore report four complementary measures.

- AUROC: the chance that a randomly selected event receives a higher risk score than a randomly selected non-event.
- Average precision: precision averaged as the ranking moves from the highest-risk listings downward; the event rate is its practical reference level.
- Brier score: mean squared probability error, which rewards useful calibration and is better when smaller.
- Top-20% targeting: the fraction of events captured and the precision/lift if inspectors could prioritize one fifth of listings.

$$Brier = (1/n) \sum_i (\hat{p}_i - Y_i)^2$$

Conditional performance intervals come from 1,500 address-cluster bootstrap samples of the fixed, averaged out-of-fold predictions. Differences between metadata-plus-text and its matched review-history baseline use 2,000 paired address bootstraps, preserving the fact that both models predict the same restaurants. These intervals reflect address sampling conditional on the fitted predictions; they do not refit models and therefore omit training- and split-selection uncertainty.

3.6 Stress tests

Sensitivity checks change the minimum earlier-review count to 1, 5, or 10; require at least 50% food-context reviews; exclude ungraded Z records; impose seven- and thirty-day review embargoes; and use score greater than 14. A forward test trains on inspections before February 1, 2015, and tests on later inspections. Finally, an A-only ridge regression predicts the continuous 0–13 score. These are stress tests, not a buffet from which to select the prettiest number.

3.7 The two leakage questions

Temporal leakage and evaluation leakage are different. Temporal leakage asks whether later reviews enter the predictor. Evaluation leakage asks whether a model is judged on the same outcomes used to fit it. The first is tested by pre-only versus all-review TF-IDF. The second is tested with a 64-feature reconstruction from the released crawl-time stars, review count, price, and keyword columns, comparing same-data scoring with repeated grouped cross-validation. That reconstruction is not the authors' exact model; their final coefficient vector and complete fitting script are unavailable.

4 Results

4.1 The sample before the algorithm gets clever

The primary sample contains 406 scores from 0 through 13, 44 from 14 through 27, and 15 of 28 or more. Archived grade labels are 404 A, 22 B, five C, and 34 Z/ungraded. The median score is 10. A listing has a median of 14 earlier reviews; the median review contains about 89 words. Visible-hygiene vocabulary appears in at least one earlier review for 273 listings, while illness language appears for 110. These mentions are common enough to model, but common is not the same as discriminating.

Table 3: Repeated grouped out-of-fold performance

Model	AUROC [95% interval]	Avg. precision	Brier	Top-20% precision	Lift
Prevalence only	0.476 [0.407, 0.548]	0.122	0.111	—	—
Stars + volume	0.518 [0.435, 0.599]	0.135	0.112	12.9%	1.02×
Review-history metadata	0.488 [0.401, 0.572]	0.128	0.113	14.0%	1.10×
Metadata + lexicon	0.440 [0.356, 0.524]	0.113	0.117	12.9%	1.02×
Metadata + TF-IDF text	0.503 [0.417, 0.588]	0.134	0.113	14.0%	1.10×

Notes: N = 465; 59 events. Intervals are conditional address-cluster bootstraps of fixed out-of-fold predictions. The prevalence baseline cannot rank listings, so its targeting cells are not applicable. Lower Brier is better; event prevalence is 0.127.

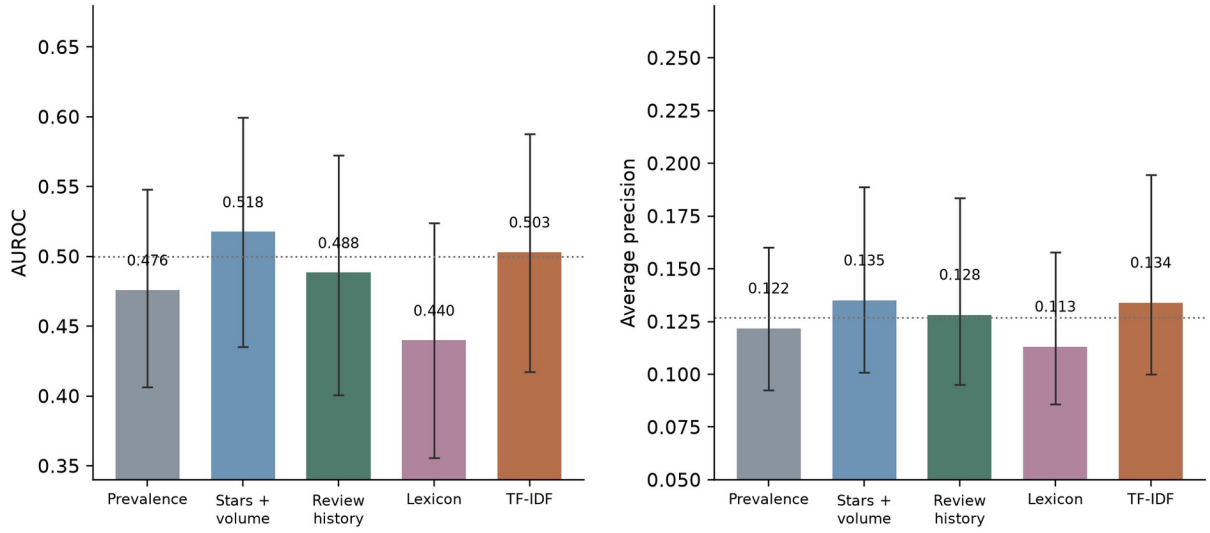

Figure 3: Out-of-fold discrimination. Error bars are conditional 95% address-cluster intervals around fixed predictions. The dotted references are 0.5 for AUROC and the 12.7% event rate for average precision.

4.2 Text adds a little—and does not earn a victory lap

The metadata-plus-TF-IDF model's AUROC is 0.503, compared with 0.488 for the matched non-text model. The paired difference is 0.015 (conditional 95% interval -0.016 to 0.043); average precision differs by 0.006. The point estimates give text a small edge over the matched baseline, but the intervals permit slight harm and do not support a stable advantage. The leaner stars-plus-volume model remains higher at AUROC 0.518. The lexicon model is worse still, with AUROC 0.440 and a Brier score above the prevalence baseline. The sanitation dictionary did not discover a cheat code.

4.3 Calibration is not impressed either

Useful risk scores should rise when observed risk rises. The stars-plus-volume model's five predicted-risk groups do not increase cleanly: observed event rates are 10.8%, 12.9%, 10.8%, 16.1%, and 12.9%. The text model is flatter still; its highest fifth has the same 14.0% observed rate as its lowest fifth. The matched review-history baseline reverses the desired order, with 16.1% in its lowest fifth and 14.0% in its highest. The lexicon's highest predicted fifth also contains fewer events than its lowest. This is not a hidden high-performance model being punished by one fussy metric. Its ranking is unstable in the raw groups.

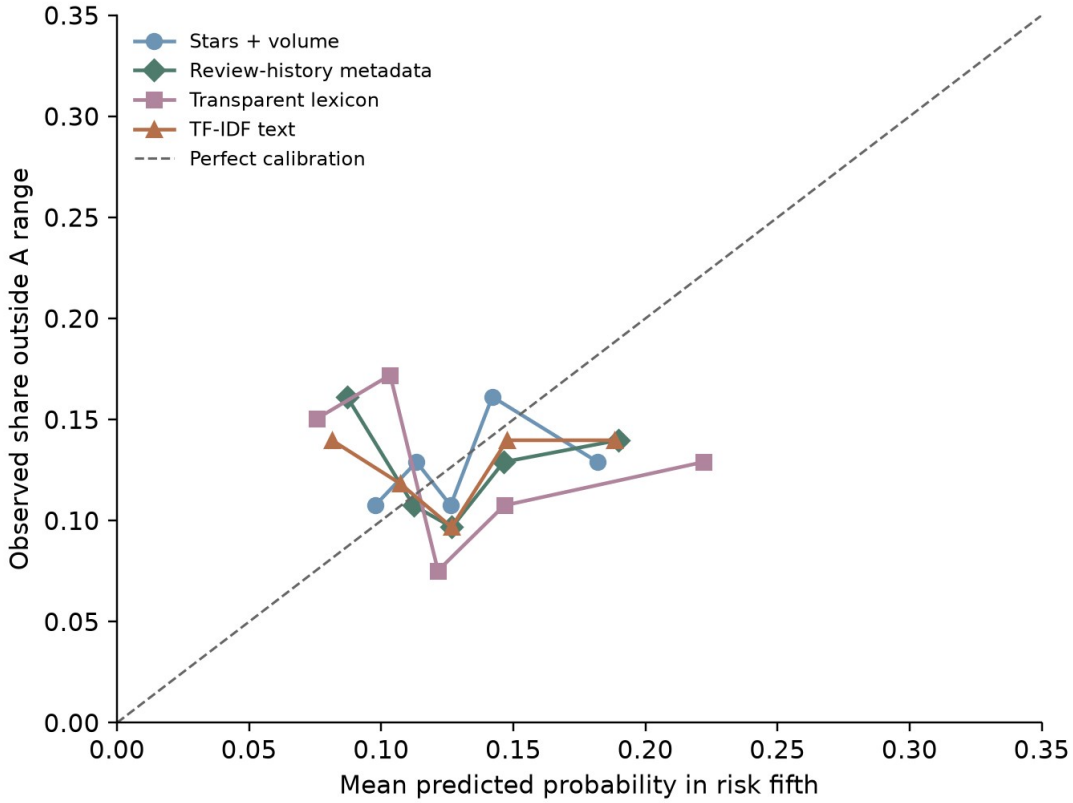

Figure 4: Calibration by fifth of out-of-fold predicted risk. A useful model should roughly follow the dashed 45-degree line and order groups from low observed risk to high.

4.4 What happens if inspectors can visit 20%?

At a one-fifth inspection budget, TF-IDF selects 93 listings and captures 13 of the 59 events. Its precision is 14.0%, compared with 12.7% prevalence, for a lift of 1.10. The matched review-history baseline also captures 13 events; stars plus volume captures 12 and has lift 1.02. Conditional bootstrap intervals for text lift run from 0.57 to 1.55, so the ranking could plausibly underperform random selection or improve it by about half. A regulator should not rearrange an inspection schedule around that uncertainty.

Table 4: Inspection event rates by earlier review themes

Theme	No mention N	No mention	Mention N	Mention
Visible-hygiene complaint	192	12.0%	273	13.2%
Illness language	355	12.4%	110	13.6%
Cleanliness praise	247	13.0%	218	12.4%

Notes: Rates are descriptive and unadjusted. A mention means at least one pre-inspection review contains a prespecified term. Wilson intervals overlap broadly for each contrast.

4.5 Tomorrow's reviews do not rescue the forecast

The all-review TF-IDF diagnostic uses 1,148 texts that arrived after inspection and 10 same-day texts in addition to the 7,480 earlier texts. Its AUROC is 0.512, only 0.009 above the strict pre-only value, and its Brier score improves by just 0.0001. This finding does not prove that future information is harmless. The archive contains Yelp's top-ranked reviews at crawl time, and only 26 of the 59 high-score listings have later reviews; the comparison is asymmetric. It does show that the main failure is not fixed by simply feeding the model more hindsight.

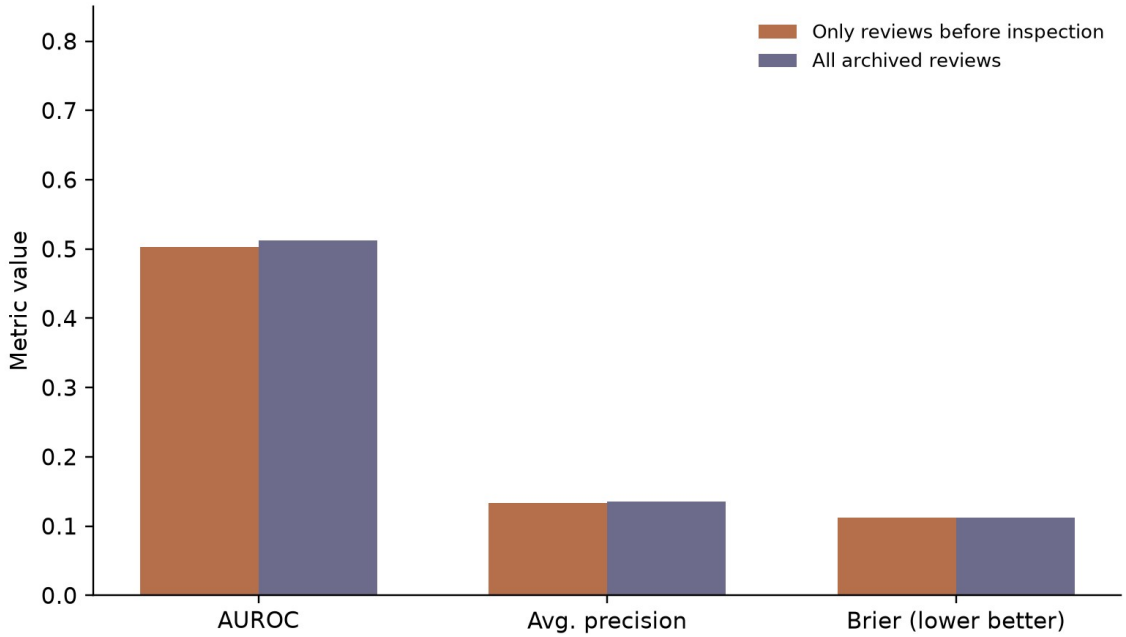

Figure 5: Pre-inspection versus all-archived text on the identical 465 listings and validation folds. Later reviews produce little change in any metric.

4.6 The same-data mirror is much more flattering

A different problem appears when the broad released feature set is allowed to grade its own homework. The reconstruction uses 61 released keyword columns plus crawl-time star rating, review count, and price. On all 732 deduplicated rows, fitting and evaluating the ridge model on the same data produces AUROC 0.711 and average precision 0.322. When the same 64 inputs are evaluated by repeated address-grouped cross-validation, AUROC falls to 0.480 and average precision to 0.122. The same-data model's apparent lift of 2.11 at a 20% budget becomes 0.97 on unseen groups.

This is not an exact reproduction of the 2016 coefficient model, and it should not be described as one. It is an audit of an evaluation choice using the released columns. The result is a general lesson: dozens of weak, correlated text features can fit idiosyncrasies in a few hundred rows. Cross-validation is not a ceremonial final step; it changes the answer.

Table 5: Released 64-feature reconstruction

Evaluation	AUROC	Avg. precision	Brier	Top-20% precision	Lift
Same-data evaluation	0.711	0.322	0.099	26.5%	2.11×
Repeated grouped CV	0.480	0.122	0.125	12.2%	0.97×

Notes: N = 732; 92 events. Same-data evaluation fits and scores the identical rows. Grouped CV learns every transformation on training folds and tests on unseen normalized addresses.

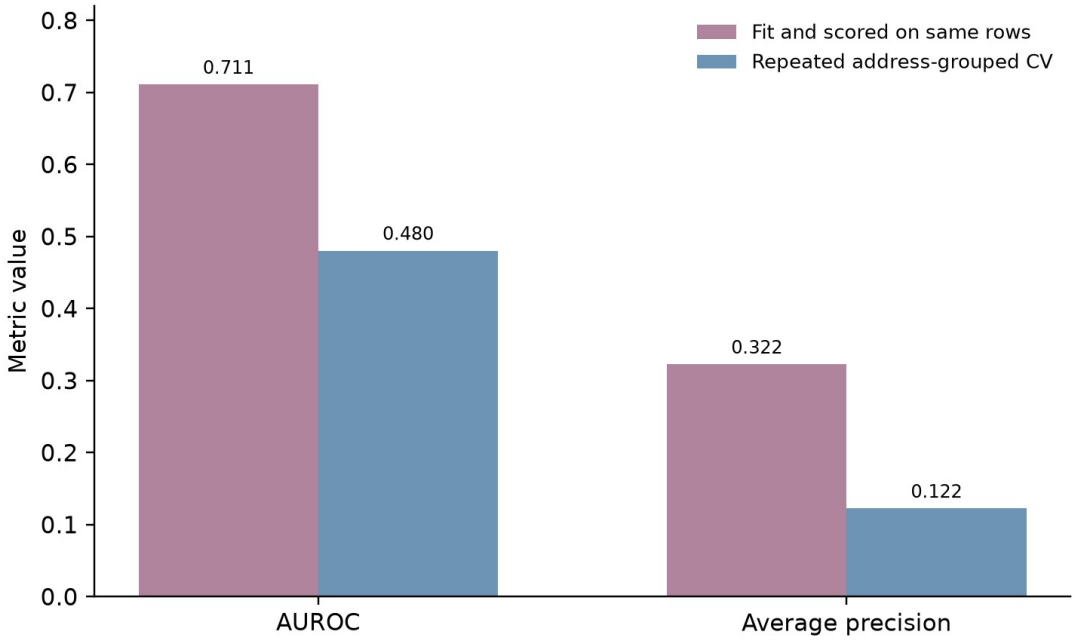

Figure 6: The optimism gap in the broad released-feature reconstruction. Same-data performance is descriptive fit, not prospective prediction.

4.7 The robot learned sauce, not sanitation

The strongest positive full-sample TF-IDF coefficients include sauce, chicken with, toast, pio, Dominican, eggs, and green. Negative coefficients include cafe, soda, pasta, and bakery. These terms are not violation mechanisms. They are likely cuisine, venue, neighborhood, or sampling proxies—and they change when the training sample changes. The coefficient plot is included precisely because it is a warning against giving colorful stories to a weak model.

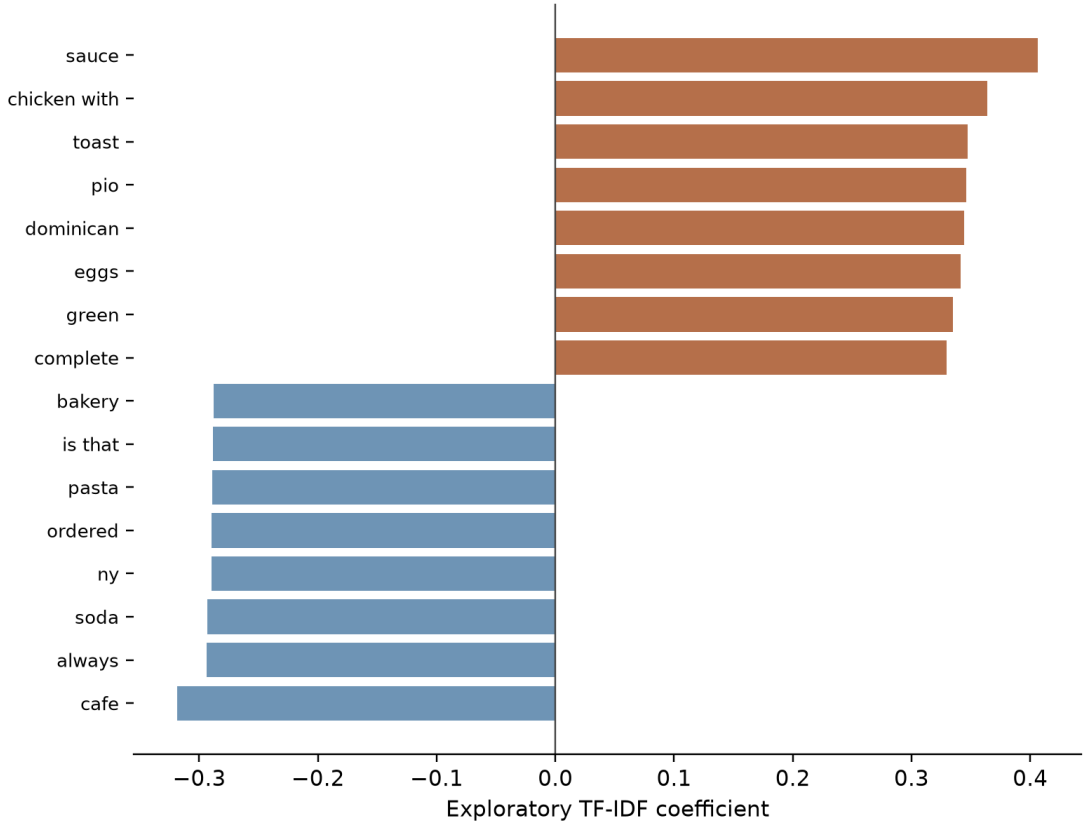

Figure 7: Largest-magnitude coefficients from an exploratory model fit to the full primary sample. They are descriptive and unstable, not causal word effects.

4.8 The later-month test is mildly better—and still modest

The forward split trains on 213 inspections through January 2015, including only 21 events, then predicts 252 later inspections with 38 events. Here TF-IDF reaches AUROC 0.571 and average precision 0.199, while the matched review-history baseline reaches 0.572 and 0.202; stars plus volume reaches 0.585 and 0.188. Text and the matched baseline each capture 10 of 38 events in the top 20%, a lift of 1.30. This single split is better than the repeated-fold average, but text does not improve on matched metadata and still trails the simple model on AUROC and probability error. The disagreement between tests is itself evidence of instability.

Table 6: Forward-in-time holdout beginning February 1, 2015

Model	AUROC	Avg. precision	Brier	Events in top 20%	Lift
Prevalence only	0.500	0.151	0.131	—	—
Stars + volume	0.585	0.188	0.129	9/38	1.17×
Review-history metadata	0.572	0.202	0.129	10/38	1.30×
Metadata + lexicon	0.546	0.176	0.135	9/38	1.17×
Metadata + TF-IDF text	0.571	0.199	0.130	10/38	1.30×

Notes: Training N = 213 (21 events); test N = 252 (38 events). The prevalence model cannot rank tied test probabilities, so targeting is not applicable. This is one chronological split, not repeated cross-validation, so its point estimates are especially sample-dependent.

4.9 Robustness: the needle moves, but it does not settle

Table 7: TF-IDF sensitivity checks

Specification	N	Events	AUROC	Avg. precision	Brier
At least 1 prior review	597	74	0.465	0.110	0.110
at least 3	465	59	0.478	0.124	0.114
At least 5 prior reviews	396	47	0.539	0.141	0.105
At least 10 prior reviews	309	34	0.549	0.125	0.099
Food-context share at least 50%	440	57	0.518	0.149	0.114
Exclude ungraded Z records	431	27	0.497	0.091	0.059
Seven-day review embargo	464	59	0.563	0.158	0.111
Thirty-day review embargo	459	58	0.514	0.137	0.112
Legacy cutoff: score greater than 14	465	56	0.528	0.133	0.107

Notes: Each row uses three repeats of five-fold address-grouped validation, so the primary row differs slightly from the five-repeat headline estimate. The range, rather than any single maximum, is the point.

The most flattering specification is the seven-day embargo (AUROC 0.563); requiring ten earlier reviews reaches 0.549, while allowing a single review falls to 0.465. But there is no monotone pattern in review history, food-context filtering, embargo length, or outcome cutoff. A signal that depends on which sensible door one enters through is not yet a sturdy signal.

These rows are sensitivity checks, not nine independent chances to choose a winner. They use overlapping restaurants, different event rates, and fewer cross-validation repeats than the headline model. The table therefore supports an instability claim; it does not turn the largest observed AUROC into a confirmed result.

4.10 All A's are not equal, but Yelp still cannot sort them

The within-A analysis keeps 404 archived A records with scores from 0 through 13. A training-fold mean predictor has out-of-sample mean absolute error (MAE) 2.59 points. Stars plus volume and the expanded review-history baseline both round to 2.60; the lexicon reaches 2.58; and TF-IDF returns to 2.59. Every predictive R^2 is approximately zero or negative. In practical terms, the models do not recover score detail hidden behind the common A label.

Table 8: Prediction of continuous inspection score among archived A records

Model	MAE	RMSE	Predictive R^2	Spearman ρ
Training mean	2.59	3.26	-0.001	-0.048
Stars + volume	2.60	3.28	-0.012	-0.276
Review-history metadata	2.59	3.28	-0.012	0.011
Metadata + lexicon	2.58	3.26	-0.000	0.125
Metadata + TF-IDF text	2.59	3.26	-0.001	0.080

Notes: N = 404. Predictions use five repeats of five-fold address-grouped validation. Lower MAE/RMSE is better; positive R^2 indicates improvement over predicting the test outcome by a constant mean.

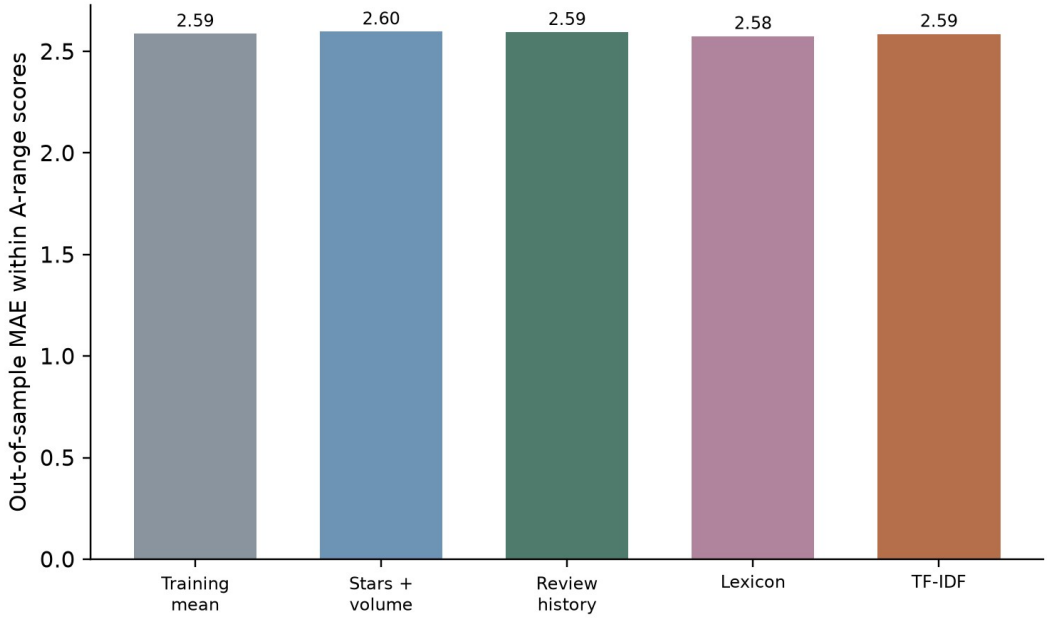

Figure 8: Within-A out-of-sample mean absolute error. Differences from the training-mean baseline are less than two hundredths of a point.

4.11 Result ledger

1. Strictly earlier text gives small point-estimate gains over matched review-history metadata, but paired intervals include no gain and the leaner stars-plus-volume model has higher AUROC.
2. Including later archived reviews barely changes the result; temporal hindsight is small in this reconstruction.
3. Scoring a broad model on its own training rows creates a large optimism gap that disappears on unseen addresses.
4. A single forward split is mildly encouraging but neither superior to simple metadata nor stable across specifications.
5. Within the A range, text does not recover meaningful point-score variation.

5 Discussion

5.1 A negative result can still answer the question

The strict answer is not that Yelp is useless. It is that this open archive does not demonstrate stable, incremental inspection prediction once review timing and unseen-address evaluation are enforced. Adding TF-IDF text to matched review-history metadata raises AUROC from 0.488 to 0.503, but the conditional paired interval crosses zero; both models capture the same 13 events in their top fifth. The text model's Brier score also remains worse than the prevalence baseline, while the leaner stars-plus-volume model reaches AUROC 0.518. Those facts are more informative than a decorative accuracy number.

This conclusion is narrower than the earlier literature. Farronato and Zervas had a much larger proprietary dataset, violation-level outcomes, and a research design able to distinguish hygiene dimensions [4]. Sadilek and colleagues evaluated social-media targeting prospectively in a different city and platform [5]. The current result says that a small, top-ranked, historical Yelp sample cannot reproduce that promise with a generic open-data model. Different question, different evidence.

5.2 Why the crowd may see dirt but miss the score

First, the constructs do not line up perfectly. Inspectors observe temperature logs, food storage, surfaces behind equipment, plumbing, pest evidence, and procedural details during a single visit. Diners observe the front-stage experience over months or years. A restaurant can look clean and still earn technical points; a diner can see a grimy bathroom on a day when the later inspection is good. Review text and inspection score are noisy windows onto related but nonidentical systems.

Second, reviews are selected observations. People choose which restaurants to visit and which experiences deserve a post. Extreme experiences may be overrepresented, and platform filtering is related to reviewer and business incentives [6]. This can make review language socially interesting while weakening its value as an unbiased surveillance sample.

Third, the page itself is a selection mechanism. The archive keeps at most 40 top-ranked reviews observed at crawl time, not every review as it would have appeared before each inspection. Even a review written in 2012 may be in the sample because of votes or ranking changes that occurred later. The temporal cutoff fixes the review's publication date, but it cannot reconstruct the old Yelp page. This is a retrospective pseudo-forecast, not a time machine.

5.3 What the optimism gap teaches

The sharpest empirical pattern is not future-review inflation; it is same-data inflation. Sixty-four predictors can separate the rows they were allowed to learn, especially when many are correlated and rare. But public policy needs performance on tomorrow's restaurant, not eloquence about yesterday's spreadsheet. The drop from 0.711 same-data AUROC to 0.480 grouped-CV AUROC is a concrete demonstration of why validation rules belong in the research question itself.

5.4 Policy interpretation

Inspection triage is not a leaderboard. A model changes who receives attention, how quickly, and with what false-alarm burden. At a 20% capacity, the text model's precision is 14.0%, meaning roughly six of seven flagged listings remain in the A-point range in this archive. Such a signal might be one small input to human review, but the current evidence does not justify replacing routine scheduling. A prospective study should freeze reviews before assignment, compare model-guided and usual-practice queues, and report both violations found and inspections spent.

5.5 Limitations

1. The matched archive is selected, historical, and small. It is not a census of 2014–2015 New York restaurants and cannot establish present-day performance.
2. The original address match cannot be fully re-audited because the file omits Yelp business names, stable business IDs, cuisine, and CAMIS identifiers. A food-context sensitivity check reduces obvious venue mismatch but cannot certify every row.
3. Inspection type and adjudication stage are absent. A score of 14 or more is therefore a point-range outcome, not proof of the placard visible to diners.
4. Top-ranked reviews were observed at crawl time. Publication dates can be filtered, but historical ranking and usefulness votes cannot be reconstructed.
5. Only 59 primary events are available. Conditional bootstrap intervals are wide and omit model-refitting uncertainty; rare phrases are unstable, and more complex neural models would add capacity faster than evidence.
6. The dictionary is intentionally blunt. Negation, sarcasm, multilingual reviews, euphemisms, and context such as 'not dirty' can be misread.
7. Reviews and inspections may measure different periods. A long pre-review history can dilute recent changes, while a single inspection is itself a noisy snapshot.
8. No causal claim follows from prediction. Restaurant practices, customer mix, neighborhood, cuisine, and platform behavior may jointly influence text and scores.

5.6 A better next study

The clean next design is prospective. With permission from Yelp and NYC DOHMH, researchers would freeze all reviews available at a fixed weekly cutoff, match them to CAMIS identifiers using name, address, phone, and coordinates, and predict only initial inspections not yet conducted. Models would be locked before outcomes arrive. A randomized pilot could compare ordinary inspection scheduling with a small review-informed supplement, while auditing performance across boroughs, cuisines, price levels, and review volumes. That study could answer the policy question this archive can only rehearse.

6 Conclusion

This study rebuilt 7,480 Yelp reviews so that every word and star preceded its matched New York City inspection. It compared simple platform signals, matched review-history metadata, an auditable sanitation dictionary, and a regularized 3,000-feature text model using repeated address-grouped validation. The result is straightforward: adding words produced only small, uncertain gains over matched non-text inputs, while the leaner stars-plus-volume model retained the higher AUROC and no model reliably beat the event-rate baseline.

The study also separated two ways a model can look smarter than it is. Later reviews made little difference here. Evaluating a broad model on the rows it learned made a large difference: AUROC fell from 0.711 on the same data to 0.480 on unseen address groups. Within the A range, language also failed to predict the hidden point score. The evidence therefore favors caution, not a review-powered inspection shortcut.

Crowds can still notice events inspectors miss, and larger authorized datasets may support useful targeting. But this archive does not earn that conclusion under a strict test. Yelp brought plenty of adjectives. The clipboard keeps its job—for now.

Data and Code Availability

The source CSV is available from the Figshare deposit [3]. The reproducibility bundle accompanying this paper contains the parser, analysis, document builder, exact file checksum, derived listing-level features, out-of-fold predictions, tables, and figure files. It does not redistribute raw Yelp review text. Running the analysis on a byte-identical source file reproduces the reported values; PDF pagination may vary slightly by software and font installation.

AI Assistance Disclosure

Artificial intelligence tools were used solely to improve clarity, organization, and grammatical precision of the written manuscript. All research design, mathematical modeling, data analysis, code implementation, experimental validation, and scientific conclusions were developed and executed independently by the author.

References

[1] New York City Department of Health and Mental Hygiene. Letter Grading for Restaurants. City of New York, 2026. https://www.nyc.gov/site/doh/business/food-operators/letter-grading-for-restaurants.page. Accessed September 2, 2026.

[2] J. P. Schomberg, O. L. Haimson, G. R. Hayes, and H. Anton-Culver. Supplementing Public Health Inspection via Social Media. PLOS ONE, 11(3):e0152117, 2016. doi:10.1371/journal.pone.0152117.

[3] J. Schomberg, O. Haimson, G. Hayes, and H. Anton-Culver. New York City Dataset. Figshare, version 1, 2015. doi:10.6084/m9.figshare.1439770.v1. CC BY 4.0.

[4] C. Farronato and G. Zervas. Consumer Reviews and Regulation: Evidence from New York City Restaurants. Marketing Science, 45(4):752–772, 2026. doi:10.1287/mksc.2023.0088.

[5] A. Sadilek, H. Kautz, L. DiPrete, B. Labus, E. Portman, J. Teitel, and V. Silenzio. Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media. AI Magazine, 38(1):37–48, 2017. doi:10.1609/aimag.v38i1.2711.

[6] M. Luca and G. Zervas. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. Management Science, 62(12):3412–3427, 2016. doi:10.1287/mnsc.2015.2304.

A Parsing and Feature Rules

A.1 HTML alignment

The parser extracts datePublished tags, description paragraphs, and star-rating image labels from each raw HTML string. If D dates and T texts are present, the first min(D,T) pairs are retained. The first star image is checked against the archive's overall listing star and excluded; the next min(D,T) stars are assigned to reviews. The remaining image stars belong to page recommendations or widgets. Reviews with missing dates are excluded from temporal summaries. Same-day reviews are not treated as earlier.

A.2 Lexicon inventory

Table A1: Prespecified word families

Family	Included lowercased tokens
Visible hygiene	dirty, filthy, grimy, roach, cockroach, mouse/mice, rat, vermin, bug, fly, mold, bathroom, toilet, restroom, stench, smell, greasy, hair
Illness	sick, vomit/vomited/vomiting, diarrhea, nausea/nauseous, stomach, hospital, poisoning
Cleanliness praise	clean/cleaner/cleanest, spotless, sanitary, hygienic
Negative experience	awful, terrible, horrible, burnt, rotten, rude, slow, bland, disgusting, gross, cold, stale, worst
Positive experience	delicious, excellent, recommend, favorite, tasty, friendly, professional, amazing, wonderful, best, fresh

Notes: Each family is encoded as log(1 + mentions per 1,000 words) and the share of a listing's reviews containing at least one family token. Exact token matching does not resolve negation or sarcasm.

A.3 Data audit

Table A2: Deterministic audit values

Check	Value
Source filename	NYBIGDATAX.csv
Expected MD5	b74ed229bddde0a0c1459d6211c146f2
Source rows	743
Deduplicated rows	732
Unique address groups	726
Aligned dated reviews after deduplication	8948
Strictly pre-inspection reviews	7657
Post-inspection reviews	1280
Same-day reviews	11
Primary listings / events	465 / 59

Notes: The checksum must match before the analysis runs. Derived outputs omit raw review text, addresses, and the raw HTML field.

B Additional Results and Reproducibility

B.1 Paired text-versus-metadata differences

Table B1: TF-IDF minus matched review-history metadata

Metric	Difference	Conditional 95% interval	Interpretation
AUROC	0.015	[-0.016, 0.043]	Higher is better
Average precision	0.006	[-0.009, 0.023]	Higher is better
Brier score	-0.0003	[-0.0011, 0.0006]	Lower is better
Lift at 20%	0.000	[-0.454, 0.414]	Higher is better

Notes: Intervals resample normalized addresses and retain paired fixed predictions; they do not refit the models. For Brier score, a positive difference means TF-IDF is worse.

B.2 Reproducibility manifest

The analysis uses deterministic parsing and fixed random seeds. Primary cross-validation seeds begin at 1729; bootstrap seeds begin at 731; the paired-difference bootstrap uses 1984. Logistic models use L2 regularization with C = 1.0. Within-A ridge models use alpha = 10. Input text is never sent to an external language model by the analysis program. The bundle contains:

- code/run_analysis.py — parses the archive, fits models, and recreates all analytical tables and figures.
- paper/build_paper.py — builds the formatted manuscript from result files.
- data/derived/listing_features_no_text.csv — listing-level derived features without addresses or review text.
- results/oof_predictions.csv — repeated grouped out-of-fold probabilities.
- results/*.csv and data_audit.json — metric, sensitivity, calibration, and audit tables.
- figures/*.png — publication figures generated from the result files.
- download_data.sh — records the exact Figshare source URL and expected checksum.
- requirements.txt — tested package requirements.

The source archive itself is intentionally absent. This keeps the bundle compact and avoids republishing individual review text. The download script retrieves it from the cited Figshare record, after which the checksum gate verifies that the bytes match the analyzed version.

B.3 Reading the result without overselling it

The main result is a model comparison, not a theorem that reviews never matter. The open sample is compatible with small gains and small losses, and a better authorized dataset could change the estimate. What this paper establishes is narrower and reproducible: under these exact temporal, grouping, and validation rules, the flexible review-text model did not demonstrate a reliable improvement over matched non-text inputs. The honest ending is not a failure to find a story. It is the story.