Neighborhood Cleanliness and Yelp Review Ratings in Philadelphia

CIS 5450
Big Data Analytics

RESEARCH TEAM

Jessica Yang, Julie Dai ,Yukun Zhou

Research Objective

RESEARCH QUESTION

To what extent do neighborhood cleanliness indicators derived from 311 sanitation complaints help explain or predict Yelp review ratings?

GOAL

Build a review-level dataset that combines restaurant attributes and neighborhood cleanliness metrics

Value Proposition

Why this matters

- Yelp ratings shape food business perception
- Cleanliness reflects neighborhood living quality
- Environmental conditions rarely modeled in Yelp studies

Potential usefulness

Restaurants: better location understanding

Urban policy: public complaint data → spatial insight

Dataset Overview

Yelp Business

5,852 restaurants

location, categories,
review_count, price, rating

Philadelphia 311 Complaint Data

445,909 complaints

sanitation-related complaints
(2020–2025)

Yelp Review

686,169 reviews

individual review ratings

Neighborhood Shapefile

159 neighborhoods

neighborhood polygons

Final Dataset

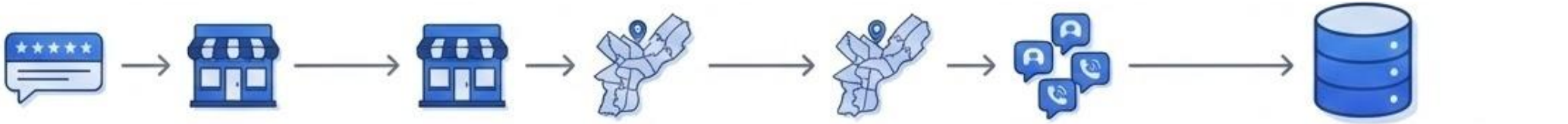
- Review-level structure: one row = one user opinion
- 686,169 review entries with restaurant attributes
- Neighborhood cleanliness metrics attached to each review
- High-granularity dataset for predictive modeling

Data Linking & Integration

Why we need it

- Datasets come from completely different sources
- No shared numeric key across Yelp, 311, and neighborhoods
- Needed a consistent way to attach cleanliness to each review

How we linked



Reviews → Businesses
(business_id)

Businesses → Neighborhoods
(spatial mapping + name cleaning)

Neighborhoods → 311
complaint aggregates

Final merge → review-level dataset with environment features

Result

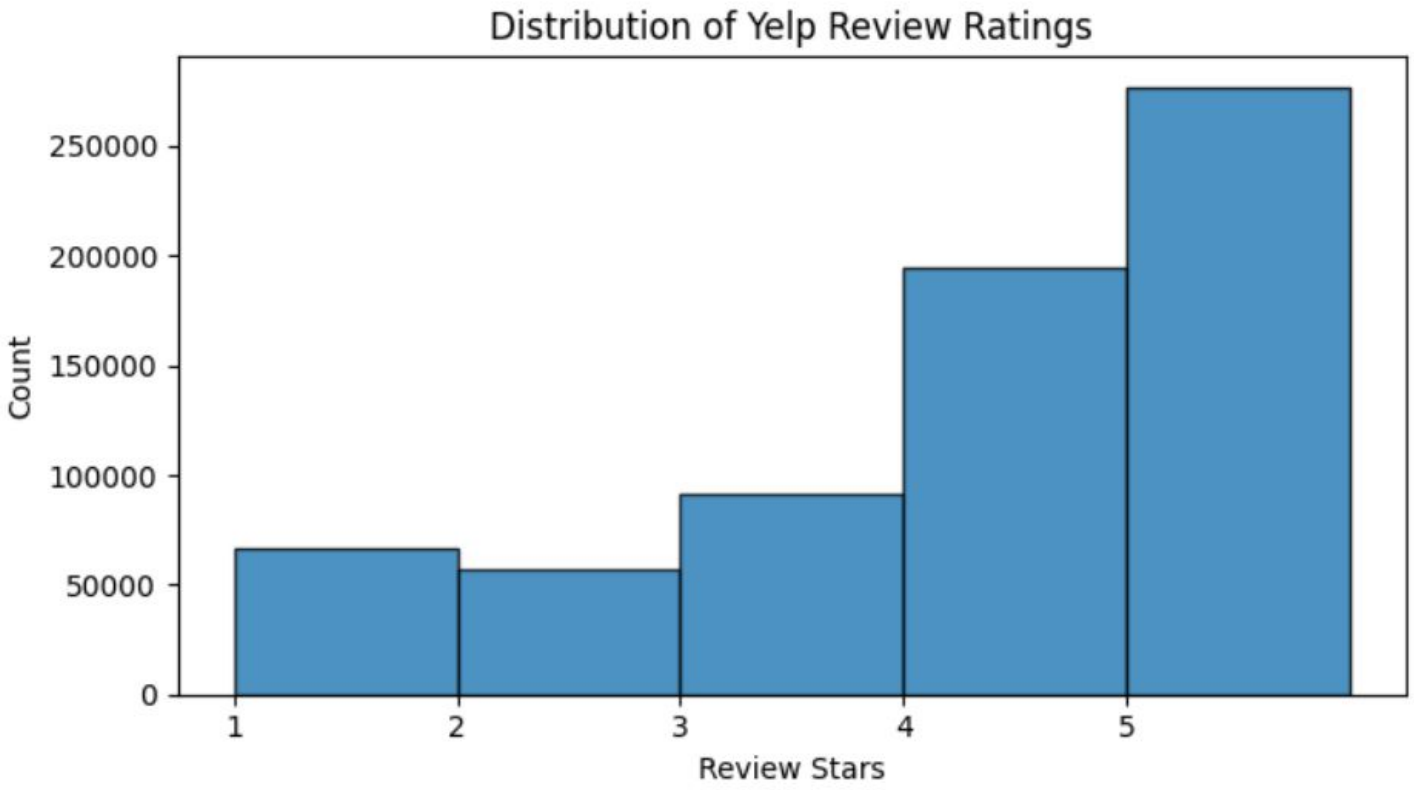
- Unified dataset connecting user opinions and neighborhood conditions

Exploratory Data Analysis: Understanding the Dataset

- Review-level dataset with Yelp + neighborhood cleanliness
- Distributions, variability, and key relationships
- Early signs of how much signal cleanliness might hold

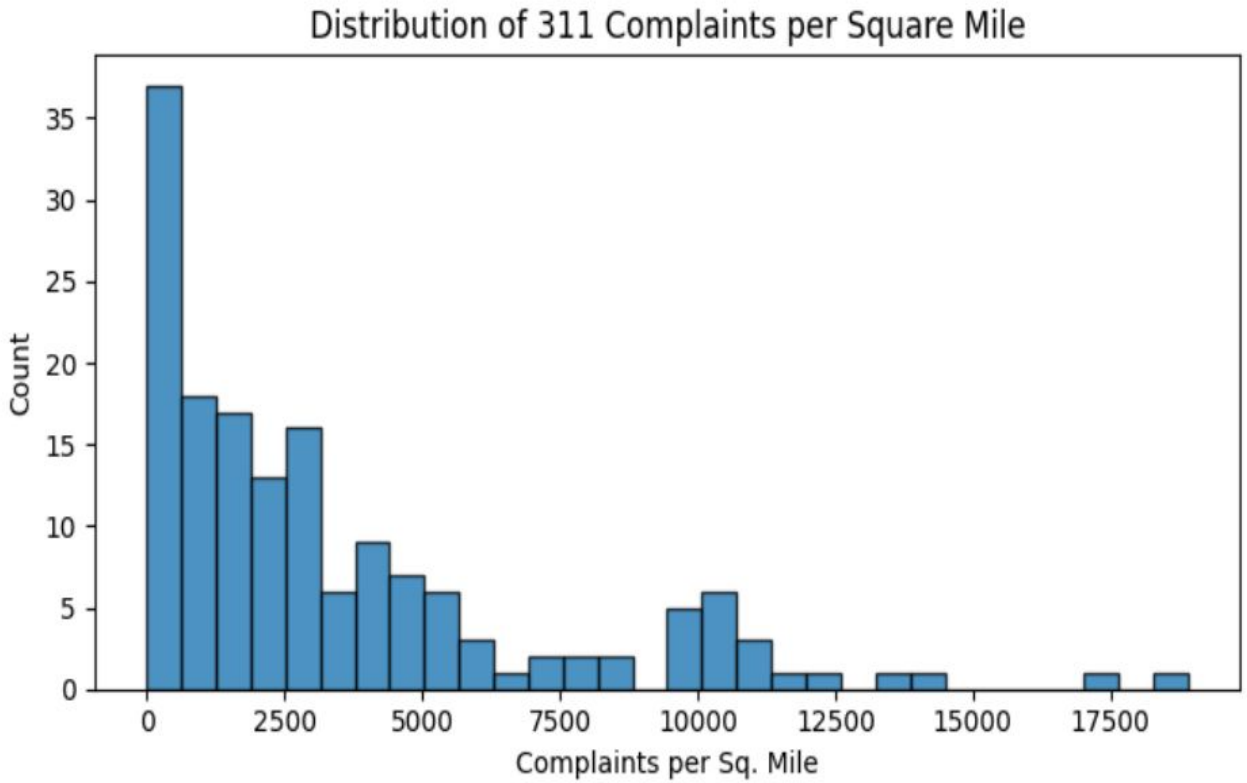
Yelp Review Ratings Are Heavily Skewed Toward High Scores

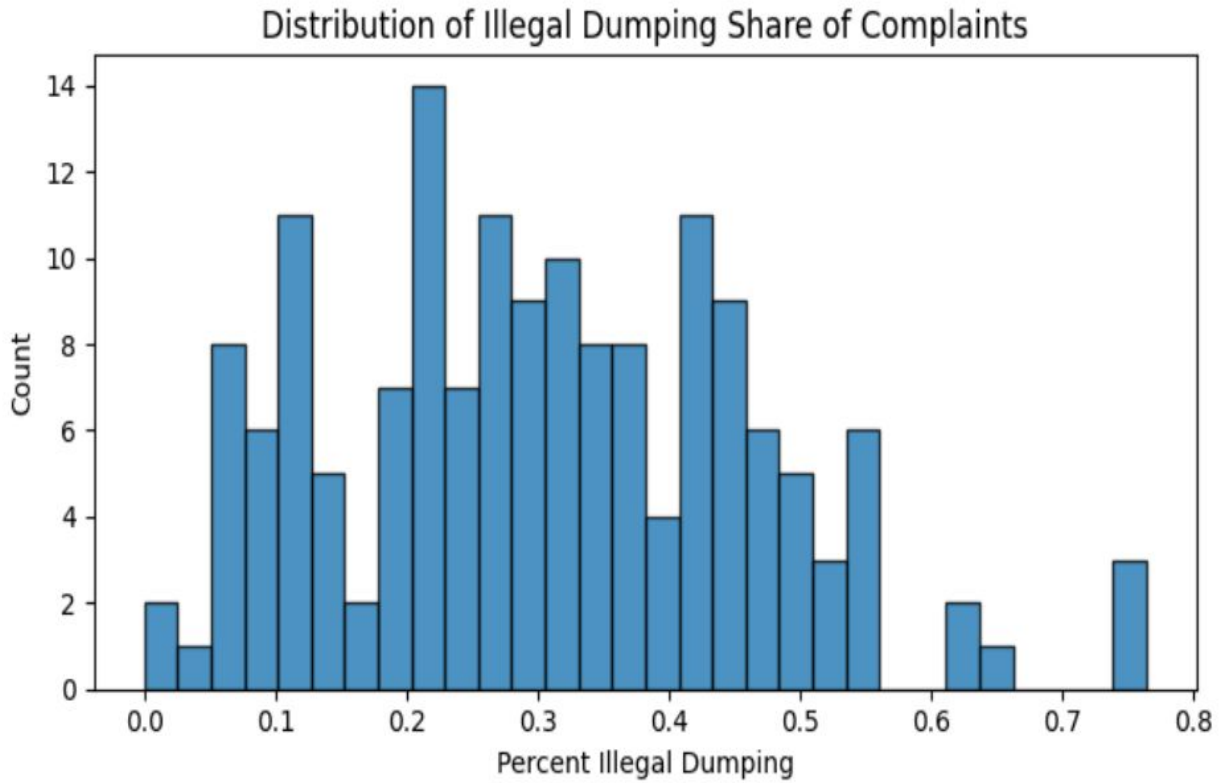
- Most reviews are 4–5 stars
- Very few 1–2 star reviews
- Predicting ratings is difficult due to low variance



Neighborhood Cleanliness Varies Widely Across Philadelphia

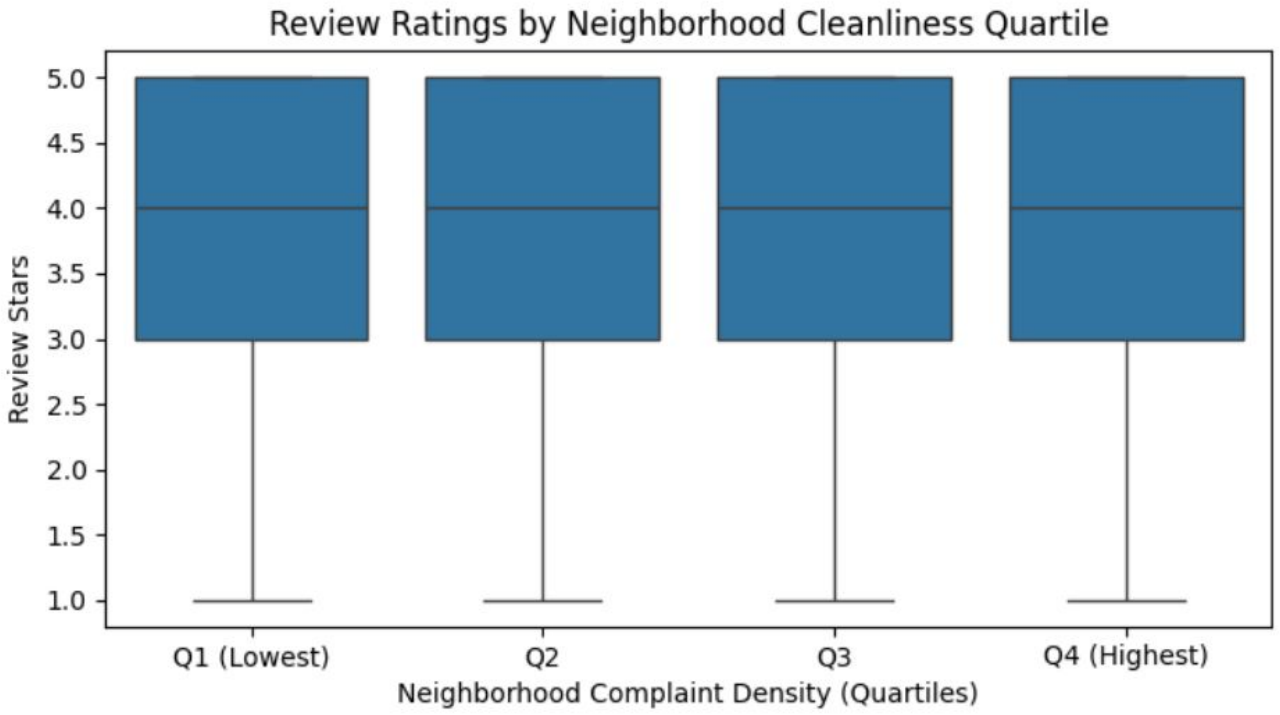
- Complaint density spans orders of magnitude
- Illegal dumping varies a lot between neighborhoods
- Cleanliness clearly has meaningful variation

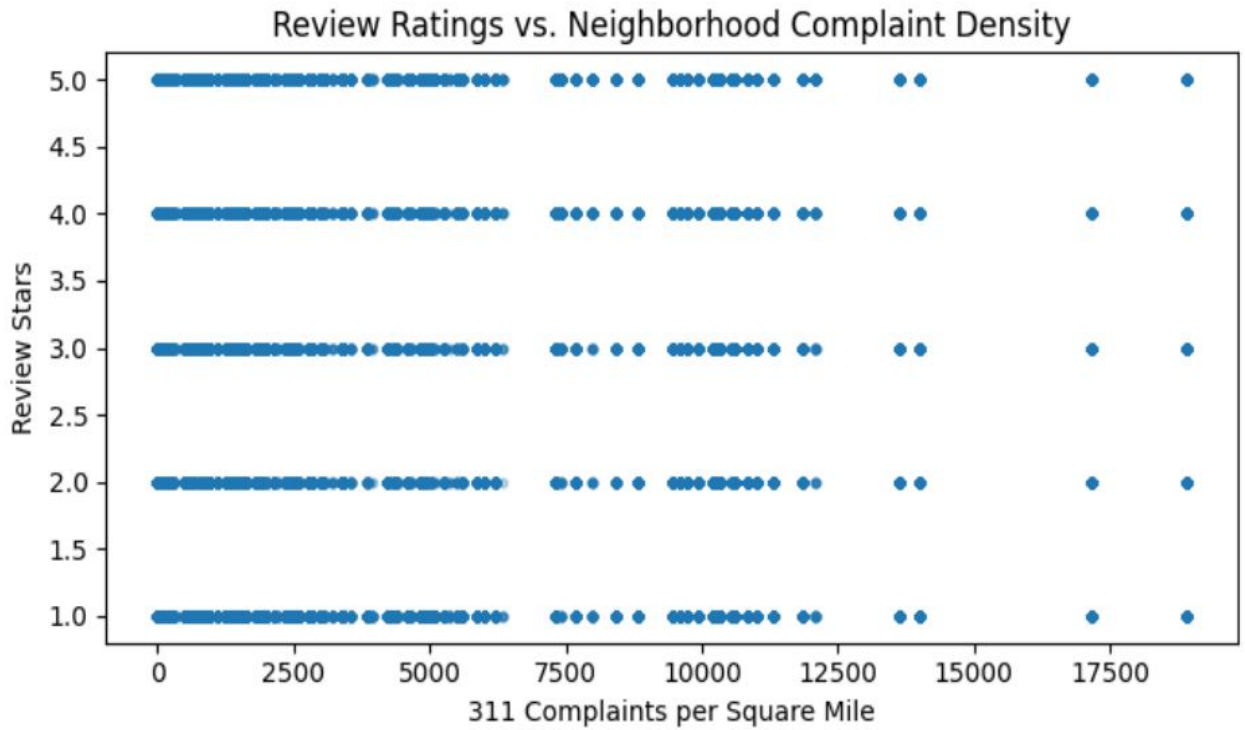




Weak Relationship Between Cleanliness and Review Ratings

- No clear downward trend
- Ratings look similar across cleanliness levels
- Early indication that cleanliness may not influence sentiment



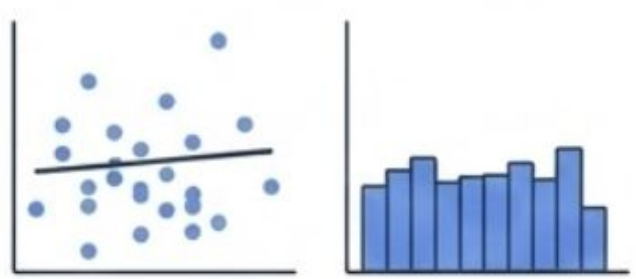


What EDA Told Us Before Modeling

Key Insights from Exploratory Analysis

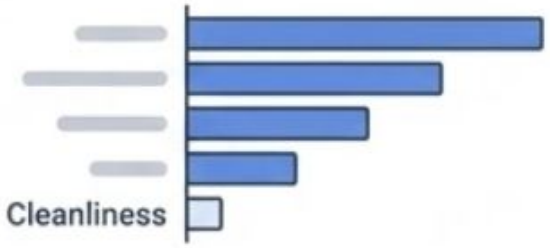
- Rating variation is limited

- Cleanliness varies, but doesn't correlate strongly



- Expect limited model performance

- Cleanliness may contribute little to prediction



Baseline Model : Yelp-Only Linear Regression

Model Overview

- Built a simple, interpretable baseline
- Uses only Yelp metadata + review engagement
- Serves as reference point before adding cleanliness features

RMSE ≈ 1.25, MAE ≈ 1.01, R^2 ≈ 0.07

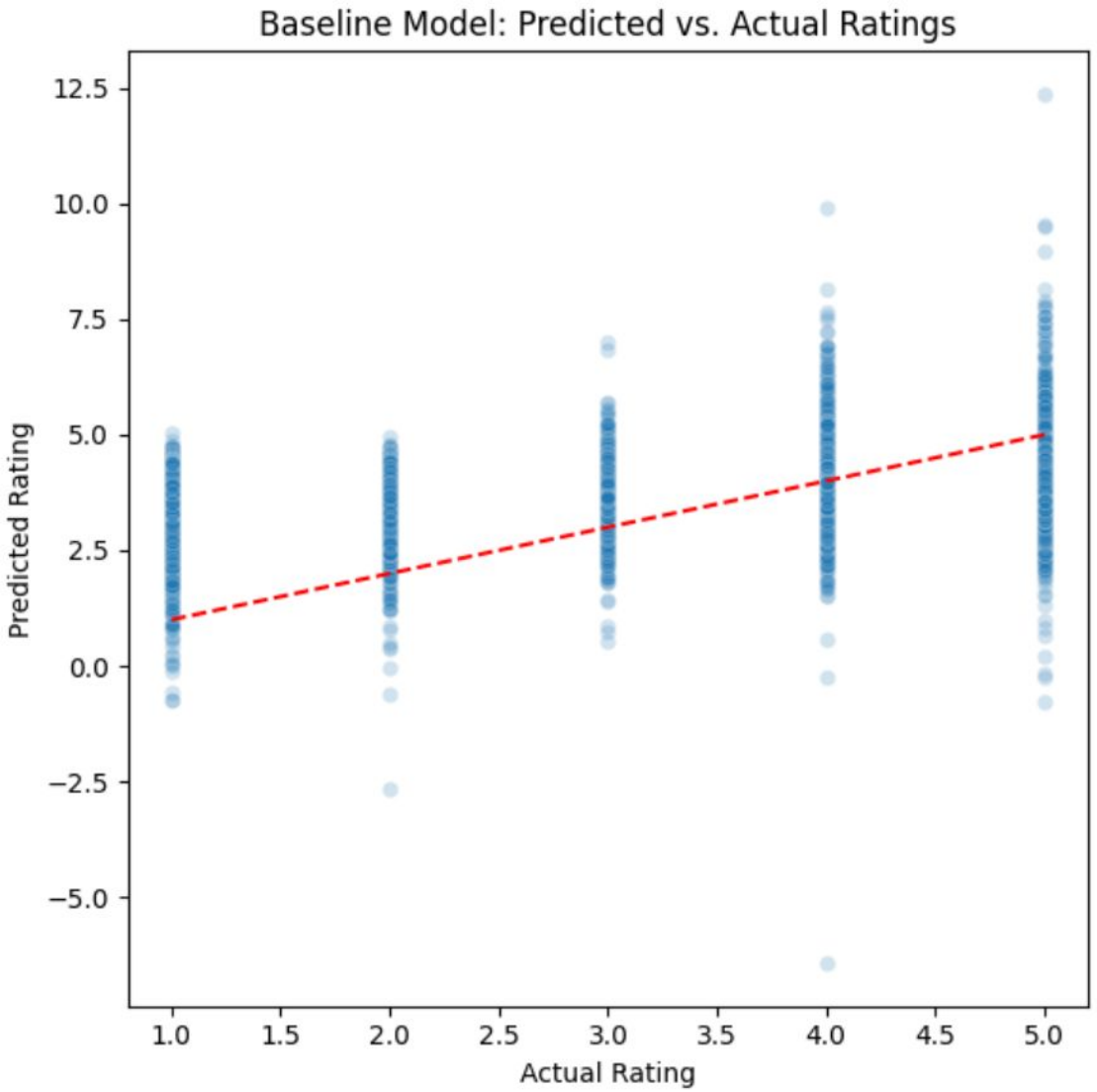
Train and test almost identical → no overfitting

Performance Metrics

Metric	Train	Test
RMSE	1.2515	1.2492
MAE	1.0127	1.0096
R^2	0.0841	0.0755

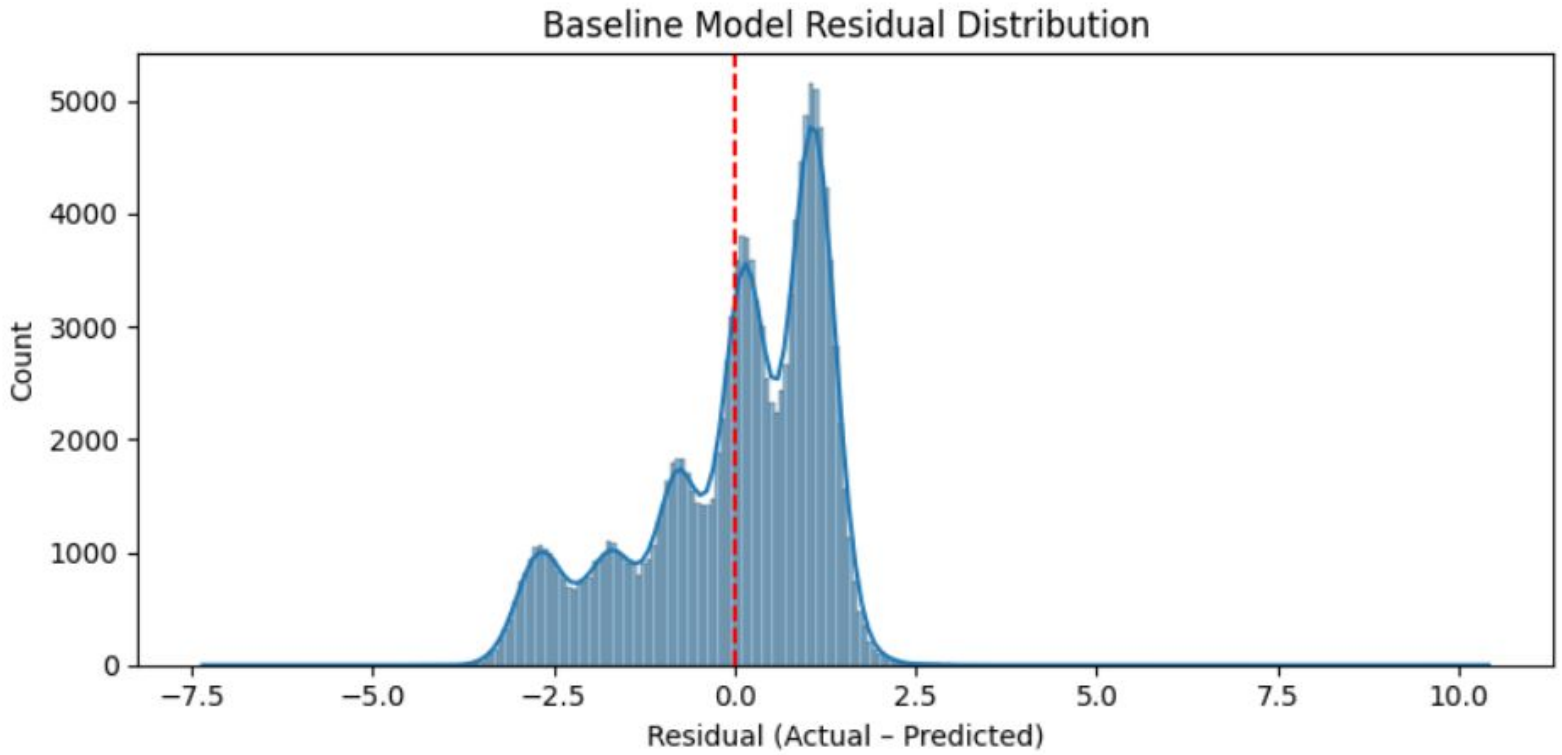
Predicted vs Actual Ratings: Weak Linear Signal

- Very wide vertical spread
- Higher ratings → slightly higher predictions
- But predictions cluster too tightly



Residuals: Mostly Within ±2 Stars

- Errors cluster between −2 and +2
- Slight left skew → mild overprediction
- Reflects difficulty of predicting exact star ratings



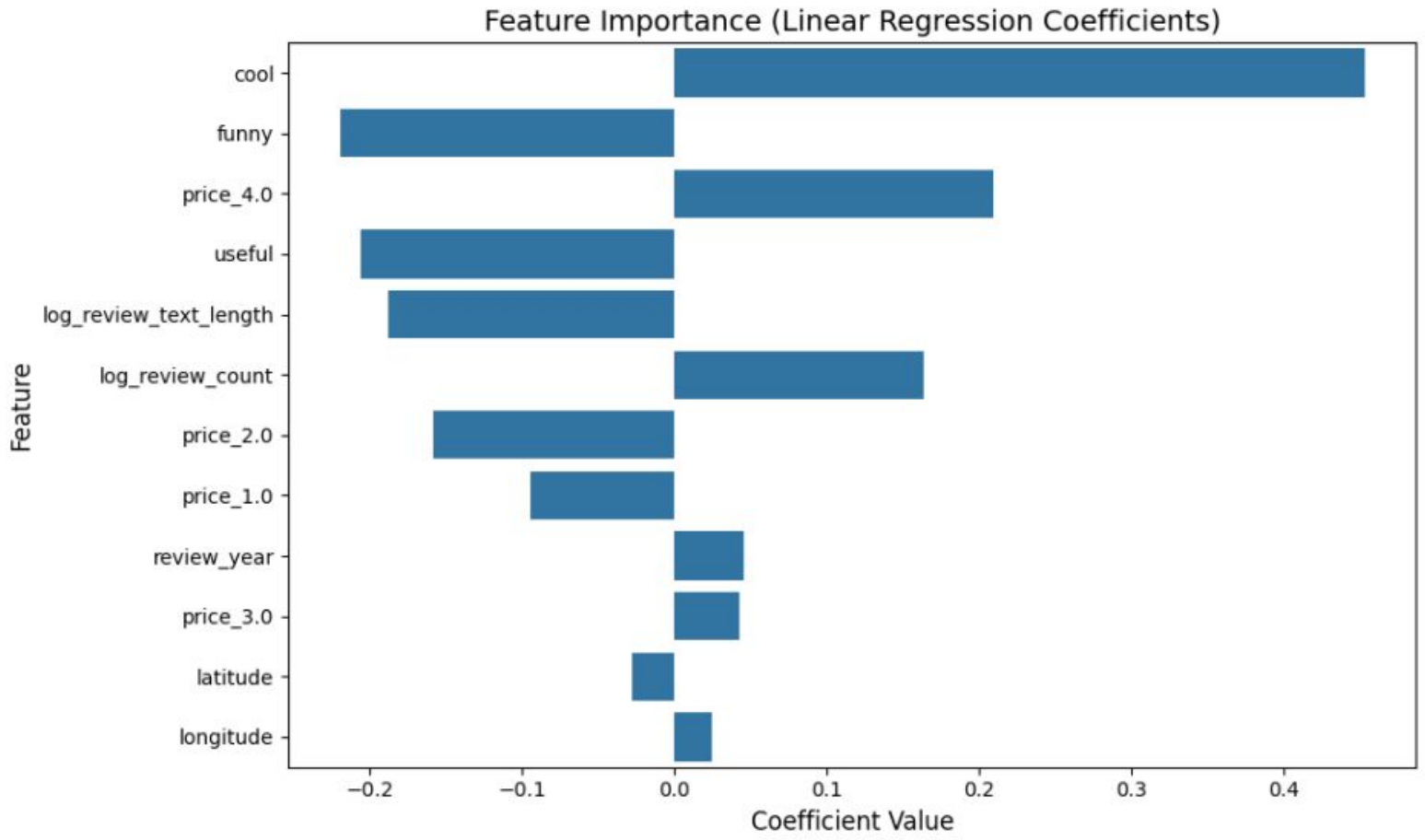
Which Yelp Features Matter Most?

'Cool' = 0.4531
strongest positive signal

'Funny' = - 0.2194 'Useful' = -0.2052
strongest negative signals

Price tier matters more than location

Text length correlates with lower ratings



Model 2: XGBoost Nonlinear Regression

Goal: Test whether adding neighborhood cleanliness improves Yelp rating prediction

Motivation: Cleanliness variables show strong correlations → nonlinear interactions likely

XGBoost Chosen because linear models cannot capture nonlinear effects

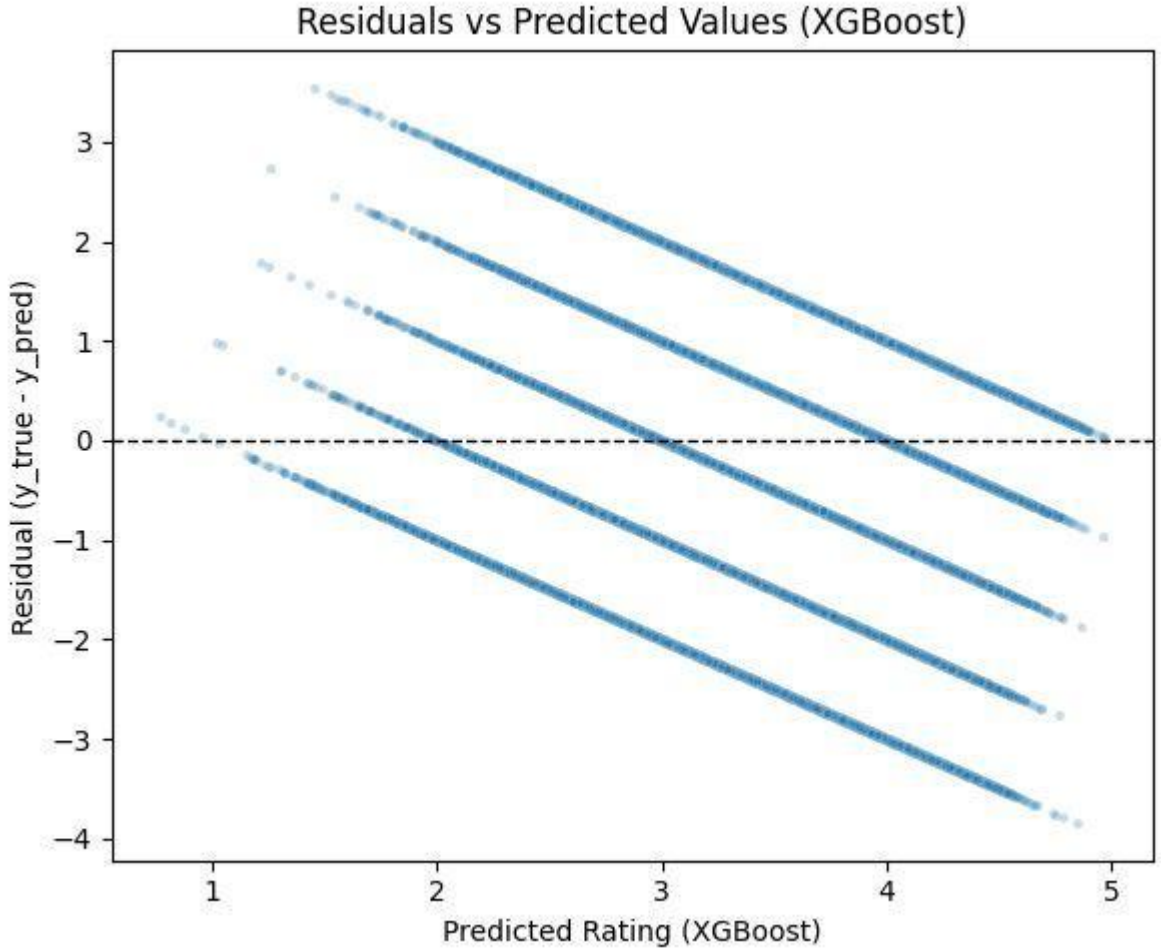
Preprocessing: Same preprocessing pipeline as baseline

Complexity Control Model complexity tuned to avoid overfitting

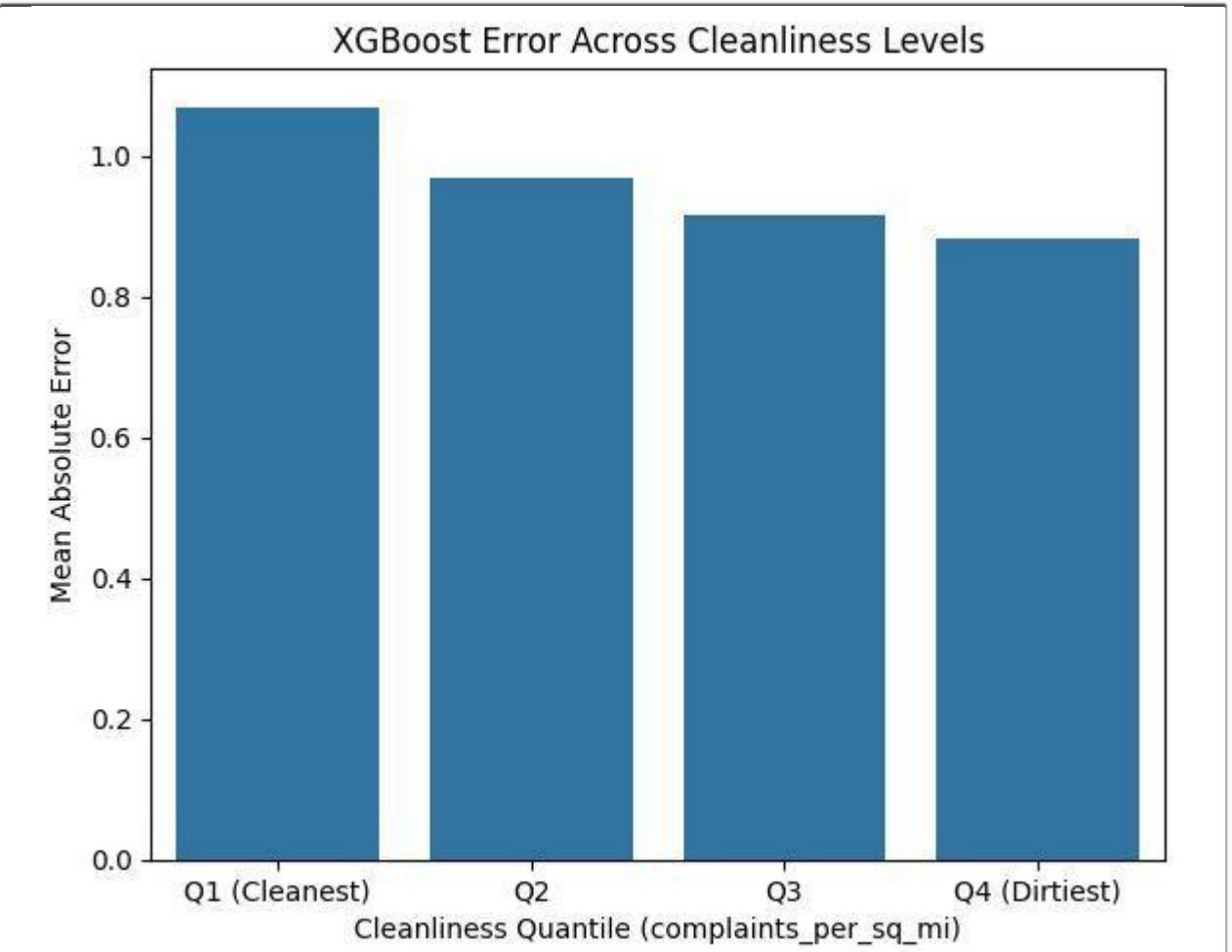
Advantage Nonlinear model better suited for Yelp + environmental signal

XGBoost Model Performance

- RMSE improved 1.25 → ~1.21
- MAE improved 1.01 → ~0.98
- R^2 improved 0.07 → ~0.13
- Improvements small but consistent across all metrics
- Meaningful given noisy and subjective Yelp ratings
- Residuals are tighter and more centered vs. baseline
- Diagonal bands arise from discrete 1–5 star ratings, not model bias



Residual Insights from XGBoost



- Much fewer extreme residuals compared to baseline
- No systematic bias across predicted rating levels
- Residual pattern confirms stable generalization
- Cleanliness contributes a small but measurable signal
- Confirms need for nonlinear modeling

Model 3: Random Forest Regression

Used as comparison nonlinear model

XGBoost: sequential boosting
Random Forest: independent trees averaged

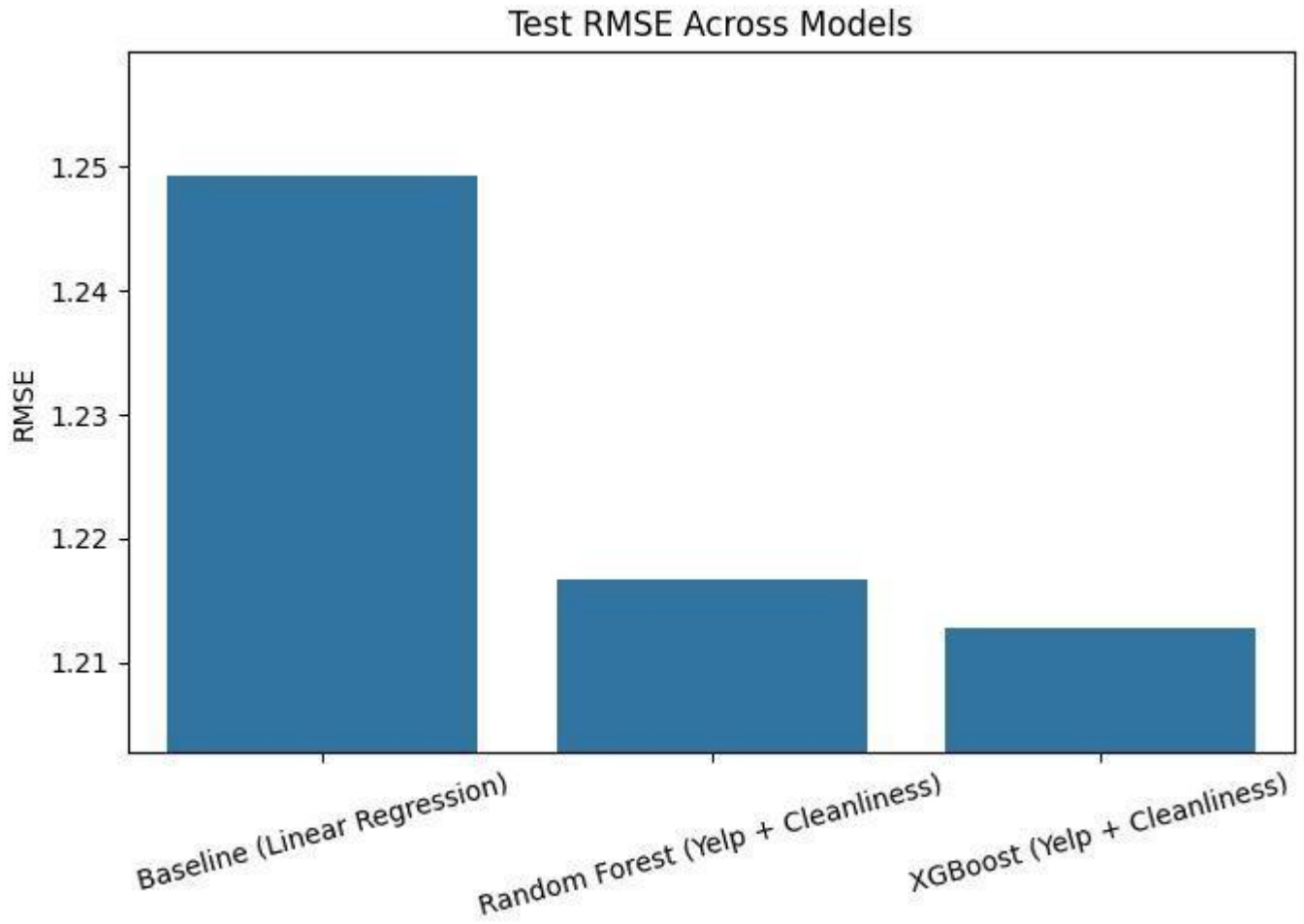
Tests whether improvements come from nonlinearity alone

Same combined feature set as XGBoost

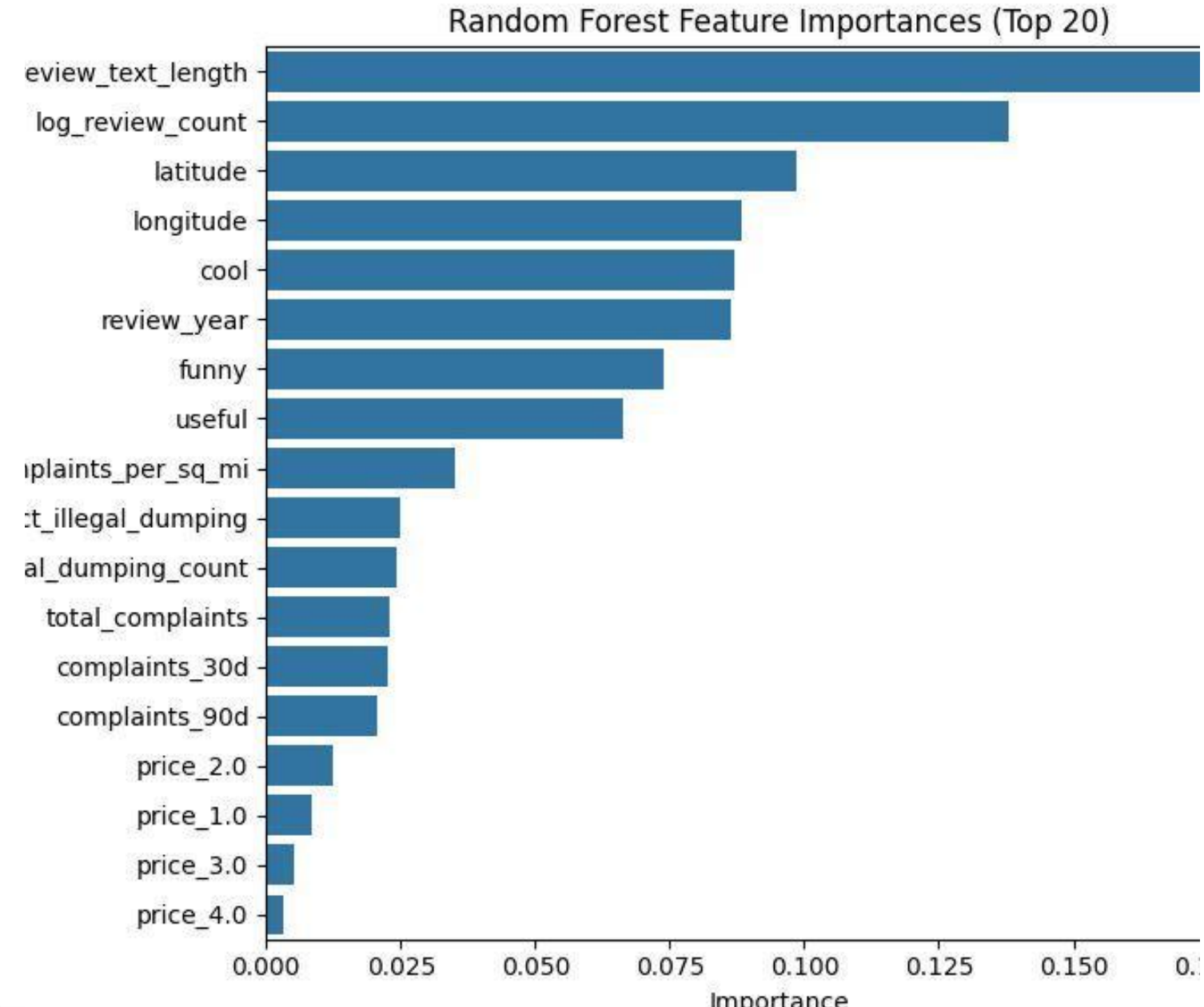
Helps validate how boosting vs. bagging behave on Yelp + cleanliness data

Random Forest Model Performance

- RMSE ≈ 1.22
  (better than baseline, slightly worse than XGBoost)
- Larger train–test gap
  → stronger overfitting than XGBoost
- Residuals more widely spread
- Confirms boosting generalizes better for this problem



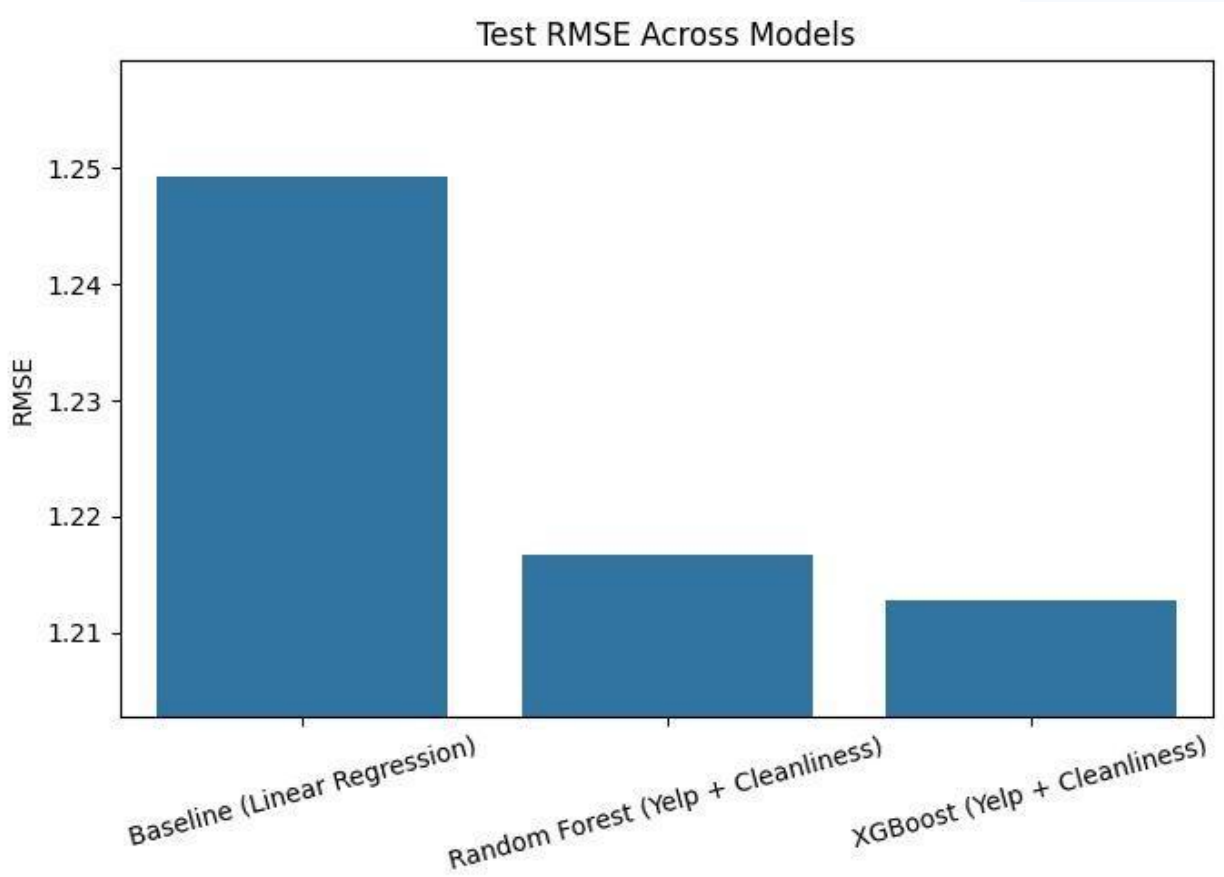
Feature Importance in Random Forest



- Yelp-based features dominate importance ranking
- Review text length & review count are top predictors
- Cleanliness features land in the mid-range
- Consistent with XGBoost interpretation
- Cleanliness matters, but isn't the primary driver of Yelp ratings

Comparing Baseline, XGBoost, and Random Forest

- Both nonlinear models outperform baseline linear model
- XGBoost provides best balance of flexibility + generalization
- Random Forest improves prediction but overfits more
- Cleanliness adds consistent but modest predictive value
- Results align with earlier EDA findings



Implications & Insights

- Yelp ratings mainly driven by restaurant-specific factors
- Cleanliness provides contextual but secondary predictive signal
- Dirtier neighborhoods → clustered ratings → easier to predict
- Cleaner neighborhoods → wider variation → higher uncertainty
- Suggests customer expectations differ by neighborhood environment
- Environmental context valuable as supplementary information

Challenges, Limitations & Future Work

Challenges

- Yelp ratings are very noisy and subjective
- Individual reviews reflect personal experiences
- Hard to predict large rating variation

Limitations

- Cleanliness features are neighborhood-level, may miss local differences
- Key restaurant factors (food quality, service) not included
- Review-level target increases noise compared to restaurant-level averages

Future Work

- Use review text (sentiment, keywords) to capture customer experience
- Create more fine-grained spatial features (street/block level)
- Add richer restaurant attributes (cuisine type, health inspections)
- Test ordered-rating models & expand to other cities/time periods

Thank You

CIS 5450
Big Data Analytics

RESEARCH TEAM

Jessica Yang, Julie Dai ,Yukun Zhou