Austin Housing Price Prediction

Ishrak Wasif Udoy — Nikhil Kumar — Mohar Chaudhuri — Pavithran Murugan

Context

In this prediction contest, our team tackled the challenge of estimating housing prices in Austin using a dataset of over 6,700 residential properties with 35 initial features. Through systematic feature engineering and model comparison, we developed a predictive model achieving $160 RMSE using 5-fold cross-validation.

Dataset Overview

Austin housing dataset with **6,784 properties** and 35 original features spanning 2019-2020. Target variable is house price (latestPrice) ranging from $5.8K to $6.25M with mean $486K. Through comprehensive feature engineering, expanded to **132 features** incorporating spatial, temporal, and textual information.

Feature Engineering

Feature engineering represents the most critical phase of this project. Starting with 35 basic features, we systematically created 97 additional engineered features through domain expertise and data exploration. The approach focused on capturing three key aspects of housing value: **spatial relationships** (location premiums), **property characteristics** (size and quality ratios), and **market signals** (text-derived luxury indicators).

1. **Ratio Features**
   - sqft_per_bedroom (layout efficiency)
   - bath_bed_ratio (quality indicator)
   - lot_to_living_ratio (land utilization)
   - room_density (space optimization)
2. **Location Features**
   - Distance to downtown, UT campus, airport
   - Proximity scores (normalized distances)
   - Cardinal directions from downtown
   - Geographic density analysis
3. **Temporal Features**
   - House age and age categories
   - Condition flags (new/old, extreme sizes)
   - Recent construction indicators
4. **Text Mining Features (75 total)**
   - Luxury keywords (granite, marble, custom)
   - Condition indicators (move-in ready)
   - TF-IDF with dimensionality reduction
   - Sentiment analysis ratios
5. **School & Quality Features**
   - School rating tiers (Excellent, Above Avg)
   - Teacher-student ratios
   - Property condition indicators
   - Size categories and luxury scores

Austin-Specific Features: Created location features specifically relevant to Austin market including distance to UT campus (major economic driver), downtown proximity (urban premium), and neighborhood clustering analysis. Text mining revealed luxury keywords like “granite”, “marble”, and “custom” significantly impact valuations.

Model Comparison & Results

Validation Strategy: 5-fold Cross-Validation

Rank	Model	CV RMSE	±Std
1	**CatBoost**	**$160**	**±$29**
2	HistGradientBoosting	$163	±$26
3	LightGBM Optimized	$163	±$27
4	Random Forest Optimized	$163	±$28
5	Gradient Boosting Tuned	$166	±$25
6	Extra Trees Optimized	$167	±$28
7	XGBoost Optimized	$172	±$25

Model Selection Process: Comprehensive evaluation of 7 algorithms using 5-fold cross-validation ensures robust performance estimates. CatBoost emerged as optimal choice due to superior handling of mixed data types (numerical, categorical, text-derived features) and built-in overfitting protection.

CatBoost is a gradient boosting algorithm that natively handles categorical features without requiring preprocessing. The model uses ordered boosting with random permutations to compute unbiased target statistics for categorical variables, preventing overfitting. For our case, CatBoost effectively processes mixed data types including neighborhoods, property characteristics, and numerical features while maintaining strong predictive performance through its built-in regularization mechanisms.

✓ **Ordered Boosting:** Reduces overfitting vs traditional gradient boosting

✓ **Categorical Handling:** Built-in target encoding prevents data leakage

✓ **Feature selection at splits:** Trees automatically choose the ”best” feature at each node

✓ **Robust Default Parameters:** Minimal tuning required

Feature Analysis & Business Impact

Feature Importance Insights: Living area dominates prediction (11.6 importance), followed by location features. The prominence of distance-based features (downtown, UT) validates Austin’s geography-driven housing market. Engineered features like lotSize_bathrooms (3.4) demonstrate value of domain-specific feature creation.

- **Size dominates:** Living area is strongest predictor
- **Location premium:** Distance to downtown/UT crucial
- **Engineered features:** Combined metrics gained importance
- **Education matters:** School ratings impact values
- **Text mining value:** Description features contributed

Model Performance & Validation: Final CatBoost model achieved **$160 ± $29 RMSE** using 5-fold CV. Consistent performance across folds indicates robust generalization. Feature engineering contributed **97 additional features** beyond original 35, with text mining and location engineering providing substantial predictive value.

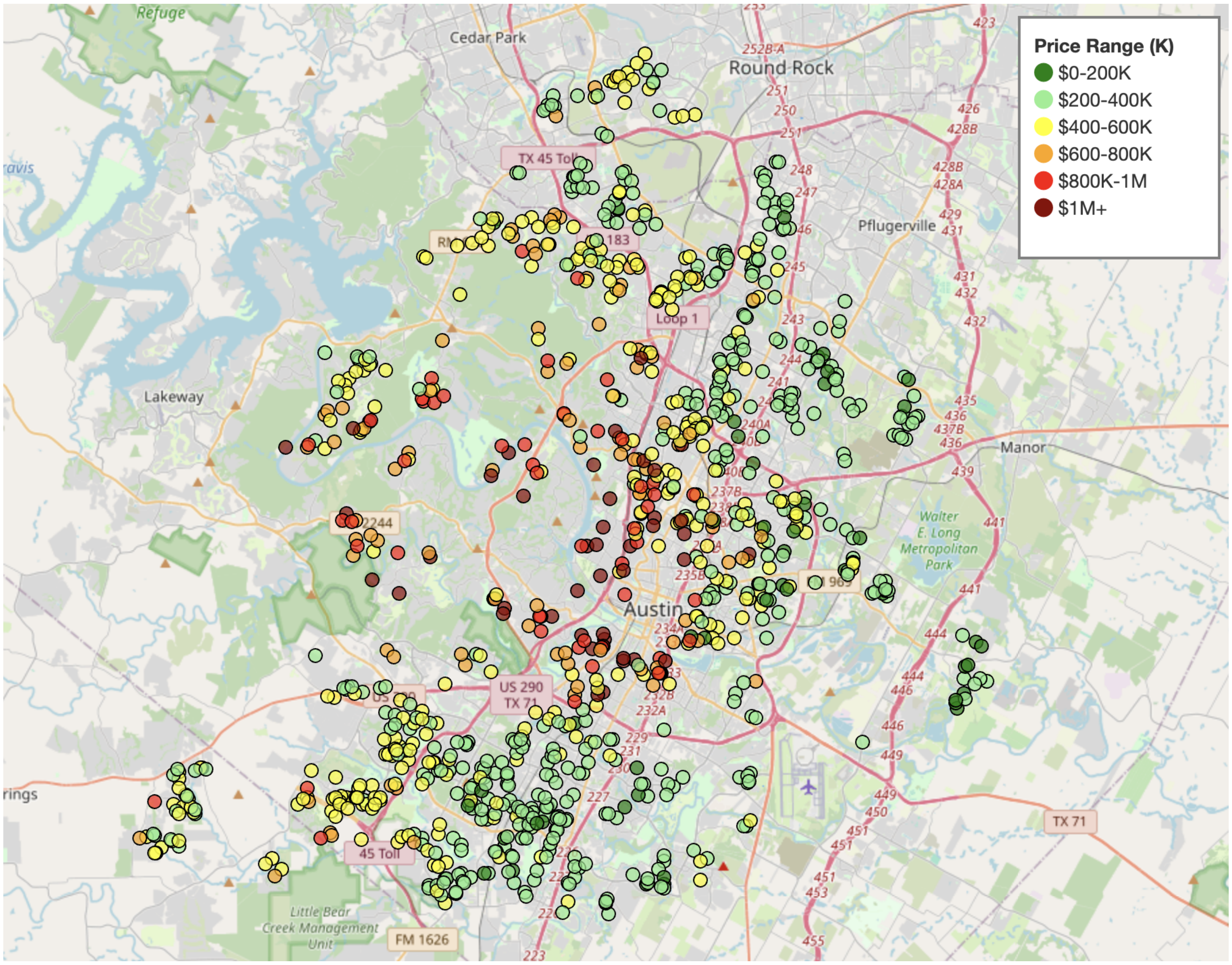

Figure 1: Excerpt from the interactive folium map of housing prices in Austin