From Coffee Shops to Housing Costs: Predictive Modeling of Neighborhood Gentrification Using Random Forest Classification

Esther Yu, Valerie Fernandez, Anisha Sawhney
Citadel Correlation-One Women's Datathon

(Dated: February 8, 2025)

This study develops a machine learning approach to predict neighborhood gentrification in U.S. metropolitan areas. Using a comprehensive dataset combining Zillow housing prices and American Community Survey demographic data from 2011-2023, we identify key socioeconomic indicators of neighborhood transformation. Our methodology employs Random Forest classification with Recursive Feature Elimination and Cross-Validation (RFECV), which identified five optimal predictive features: unemployment rate, median income, college education rate, median house value, and poverty levels. The model achieves 89% overall accuracy, with particularly strong performance in identifying non-gentrifying areas (92% precision) and good capability in predicting gentrifying neighborhoods (82% recall). This performance suggests the model's utility as an early warning system for neighborhood change. The analysis generates specific predictions for future gentrification across different ZIP Code Tabulation Areas (ZCTAs) through 2028, providing actionable insights for urban planners, policymakers, community organizations, and investors. Our findings contribute to both the theoretical understanding of urban transformation and practical tools for stakeholder decision-making. The high predictive accuracy of a parsimonious set of features suggests that gentrification follows identifiable patterns, even as it manifests differently across various urban contexts. This work provides a foundation for proactive urban policy and community planning in the face of neighborhood change.

I. Introduction

A specialized cafe replaces a family owned bottega, Luxury condos rise where modest apartments once stood. Art galleries emerge from industrial spaces. These are all signs of the transformation that urban neighborhoods across the US have experienced in the past few decades—gentrification. While some celebrate this phenomenon as an indicator of economic growth and urban renewal, others see this as harbingers of displacement and cultural erosion. As property values surge and neighborhoods transform, long-term residents—often from minority and working-class communities—find themselves priced out of homes they've inhabited for generations. Families who have weathered decades of disinvestment suddenly face impossible choices between unaffordable rents and leaving the communities they helped build. Regardless of whether this transformation represents progress or loss, it fundamentally reshapes the social fabric of cities and brings complex changes that affect housing markets, local economies, and community dynamics.

Our research aims to answer a crucial question in urban development: Can we predict which neighborhoods will experience gentrification in the next five years based on economic and demographic indicators? By developing a predictive model for gentrification, we aim to provide stakeholders—including city planners, policymakers, community organizations, and investors—with tools to better understand and prepare for neighborhood changes before they occur.

II. Methods

A. Datasets Used

Zillow Home Value Index

The Zillow data set provides monthly median home value estimates at the ZIP code level from 1996 to 2024. This data set covers all types of residential property, including single-family homes, condominiums, and cooperative housing. The home value estimates are derived from Zillow's proprietary valuation model, providing a consistent methodology across different geographic areas and time periods.

American Community Survey (ACS) 5-Year Estimates

The American Community Survey is a nationwide survey conducted by the U.S. Census Bureau that provides vital information about social, economic, housing, and demographic characteristics across the United States. For our analysis, we use the 5-year data, which pools together 60 months of collected responses to provide reliable statistics for smaller geographic areas. For example, the 2015-2019 ACS 5-year estimates represent data collected from January 1, 2015 through December 31, 2019.

The ACS dataset provides demographic and economic indicators for our gentrification analysis, including:

- Educational attainment levels
- Median household income
- Population demographics

- Employment statistics

The pooled nature of 5-year data means the statistics represent characteristics across the entire collection period rather than a specific point in time.

Geographic Concordance Data

To ensure accurate geographic matching across datasets, we employed two key concordance files. The first is a ZIP Code to ZCTA Crosswalk, sourced from a public GitHub repository. This crosswalk provides essential mapping between ZIP codes (used in Zillow data) and ZIP Code Tabulation Areas (ZCTAs, used in Census Bureau data), allowing us to accurately align Zillow housing data with ACS demographic information. [1]

Additionally, we utilized a ZCTA to County Crosswalk obtained from the U.S. Census Bureau. This dataset enables us to aggregate data at the county level when needed and facilitates our analysis of broader geographic patterns and trends. Together, these concordance files ensure consistent geographic identification across our various data sources. [2]

B. Data Cleaning

Zillow

First, we restricted our analysis to data from 2005 onwards to align with the availability of ACS data. We noted that ZIP code data was only available from 2008, which formed our effective start date. From the initial 6.7 million records, we removed approximately 480,000 entries that lacked corresponding ZCTA groupings. This reduction was deemed acceptable as these areas primarily represented regions outside our focus on metropolitan areas with populations over 65,000.

We streamlined the Zillow dataset to retain only essential columns: ZIP code, date, and median home value. We then incorporated ZCTA classifications for geographic consistency. For sporadic missing date values within a series, we interpolated using surrounding values, as these would be aggregated into annual averages in our final analysis. To align with the ACS data structure, we created 5-year rolling averages per ZCTA. For example, the 2020 value represents the average from 2015 to 2020.

Our final cleaned Zillow dataset covers 28,000 of the 33,000 total ZCTAs in the United States. The missing ZCTAs likely represent rural areas with limited housing market data or areas not covered by Zillow's tracking system.

When aligning Zillow data with the ACS 5-year structure, we initially considered using linear regression to predict housing values in a manner similar to how ACS creates their 5-year estimates. However, our analysis revealed significant heteroskedasticity in the residuals when attempting such predictions, where the variance of residuals varied systematically with the predicted values. (Figure 1) This finding makes sense given that housing prices are inherently variable and follow nonlinear patterns.

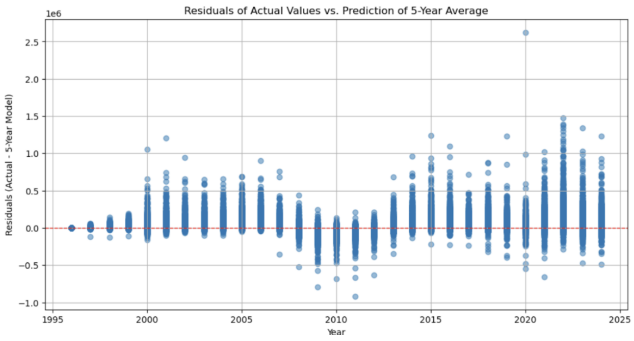

FIG. 1: Residual plot showing the difference between actual housing values and predicted 5-year averages over time. The increasing spread of residuals (fan-shaped pattern) demonstrates heteroskedasticity, indicating that housing price variations become more volatile and less predictable in recent years. This pattern supports our decision to use rolling averages rather than linear predictions for the 5-year housing value estimates.

Therefore, to align Zillow data with the ACS 5-year structure, we implemented a 5-year rolling average approach. This decision was driven by several key considerations. Figure 2 illustrates this approach using ZIP code 60637 as an example, showing how the 5-year rolling average (red line) smooths out both monthly fluctuations (blue line) and yearly averages (orange line).

We chose the 5-year rolling average method for several reasons:

1. When analyzing long-term trends and percentage increases in home values, smoother data provides more reliable indicators of sustained neighborhood change rather than short-term market fluctuations
2. This approach creates consistency with the ACS data structure, allowing for more meaningful comparisons between housing prices and demographic indicators
3. The smoothed data better captures the gradual nature of neighborhood transformation, which typically occurs over multiple years rather than in sudden jumps

ASC Data

1. **Initial Data Selection**
   We processed two main types of ACS datasets:

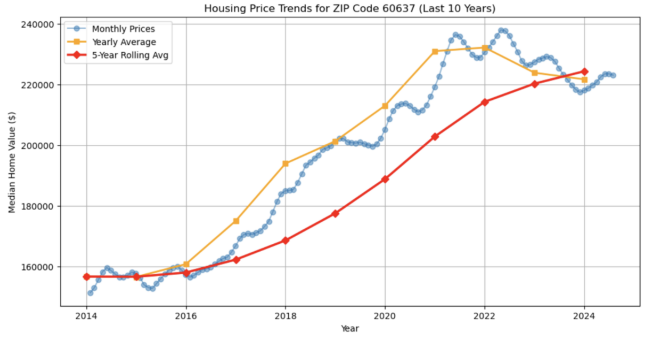

FIG. 2: Housing price trends showing monthly prices (blue), yearly averages (orange), and 5-year rolling averages (red) for ZIP code 60637. The 5-year rolling average provides a smoother trend line that better aligns with long-term neighborhood changes.

DP02 (Social Characteristics) and DP03 (Economic Characteristics). The data covered years from 2011 to 2023. For each dataset, we retained only the geographic identifiers (area name and code) and percentage estimate columns (those with suffix "PE"). This initial selection helped focus our analysis on standardized percentage metrics across different geographic areas.

2. **Geographic Area Processing**
   We extracted 5-digit ZCTA codes from the `GEO_ID` column to ensure consistent geographic identification. We made a deliberate choice to exclude Puerto Rico ZCTAs (ranges $00600 - 00999$) due to incomplete data coverage. This geographic processing ensured our analysis maintained consistency with our focus on continental U.S. metropolitan areas.

3. **Column Standardization**
   Through careful analysis of the data, we identified and removed columns that weren't consistently collected across all years. For the DP02 dataset, we removed seven specific demographic indicators that were inconsistently available. In the DP03 dataset, we removed sixteen economic indicators, primarily related to benefits data. This standardization step was crucial to ensure we only used variables that were consistently measured across our entire study period.

4. **Income Data Processing**
   Income data required special handling since it couldn't be represented as percentages like our other variables. We extracted specific income columns (`DP03_0062E` and `DP03_0063E`) and applied min-max normalization to scale the income values to a 1-100 range. This normalized income data was then reintegrated with the main dataset, ensuring compatibility with our other percentage-based metrics.

C. Gentrification Classification

For the purposes of predictive modeling, we establish clear, quantifiable criteria to classify a ZCTA as gentrifying vs. non-gentrifying over 5 years. The multi-criteria classification system requires an area to satisfy two out of the three quantitative thresholds.

1. **Rapid Housing Price Appreciation**
   A ZCTA qualifies under this criterion if its median home value increase exceeds 80% of other ZCTAs within the same county over a five-year period. This metric captures exceptional market pressure and increased housing demand relative to the broader metropolitan context.

2. **Significant Income Shift**
   This criterion is met when the area's median household income growth exceeds that of 80% of other ZCTA's in the same country over five years. This threshold identifies areas experiencing substantial economic demographic change.

3. **Significant Education Shift**
   A neighborhood meets this criterion if the percentage of residents holding college degrees increases by 10 or more percentage points over the five-year period. This substantial shift in educational attainment suggests an influx of highly educated residents.

The requirement that an area must meet at least two of these three criteria helps prevent misclassification of different types of neighborhood change. For instance, neighborhoods undergoing economic growth without displacement typically won't meet multiple criteria. This flexibility also accounts for regional variations in gentrification patterns For example, area of LA that gentrifying due to influencers may exhibit the first two criteria, rapid house price appreciation and income shift, while not having a significant increase in education.

We implemented this classification system using Python, processing our merged dataset of ACS demographic data and Zillow housing prices. The code iterates through each geographic area and time period, calculating these metrics relative to their metropolitan context and applying the two-out-of-three rule to generate binary gentrification classifications.

D. Feature Selection

To identify the most relevant predictors of gentrification, we employed Recursive Feature Elimination with Cross-Validation (RFECV). While ideally, this process would be performed for each year in our dataset to capture temporal variations in feature importance, time constraints led us to focus on 2019 data. We selected

2019 as our representative year as it provides recent pre-pandemic insights into gentrification patterns.

RFECV combines recursive feature elimination with cross-validation to automatically determine the optimal number of features. The process works by iteratively removing the weakest features while using cross-validation to evaluate model performance at each step.

Implementation
We started by loading our combined dataset and separating it into features (X) and target variable (y). For our target variable, we specifically focused on predicting gentrification status in 2023 ('gentrified_2023'). We handled missing values in our feature set by filling them with zeros to ensure data completeness.

We used a Random Forest Classifier as our base estimator, setting a fixed random state (42) for reproducibility. The model was integrated into the RFECV process with the following key parameters:

- Step size of 1, meaning features are eliminated one at a time
- 5-fold stratified cross-validation to ensure balanced class representation in each fold
- Accuracy as the scoring metric for evaluation
- Minimum of 5 features to retain in the final selection

The RFECV algorithm proceeded through these steps:

1. Started with the full feature set
2. Evaluated model performance using 5-fold cross-validation
3. Eliminated the least important feature at each iteration
4. Continued until reaching the minimum feature threshold (5 features)
5. Tracked cross-validation accuracy at each step
6. Identified the optimal number of features that maximized cross-validation accuracy

For example, our dataset included both mean and median household income metrics from the ACS data. These measures were highly correlated (correlation coefficients ¿ 0.9), making them redundant as predictors. RFECV identified this redundancy and typically retained only one of these income metrics—usually median income, as it's less sensitive to extreme values.

The RFECV analysis revealed that optimal model performance could be achieved with just five key features, shown in Figure 3. The cross-validation accuracy peaked with these five features, and adding more features did not improve model performance as seen in Figure 3.

The five optimal features identified were:

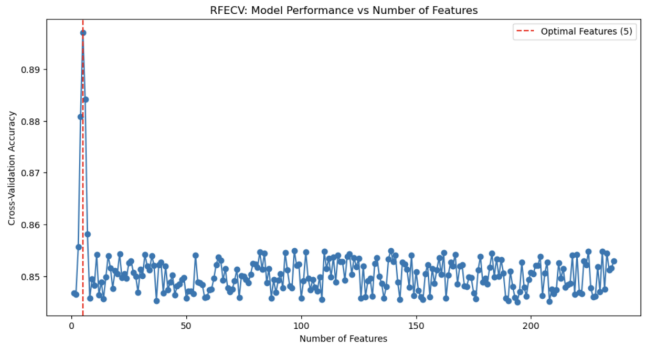

FIG. 3: RFECV performance analysis showing cross-validation accuracy versus number of features. The red dashed line indicates the optimal feature count (5), where the model achieved peak accuracy. Additional features beyond this point did not improve model performance, demonstrating that a concise set of well-chosen predictors is sufficient for gentrification prediction.

E. Model Development

While we initially considered multiple modeling approaches, including logistic regression as a baseline, we ultimately selected Random Forest as our primary classification model based on the following characteristics

1. **Handling Nonlinear Relationships**
   Gentrification is inherently a complex social process characterized by nonlinear interactions between various factors. While logistic regression assumes linear relationships between features and the target variable, Random Forest can capture complex, nonlinear patterns in the data. For example, the impact of increasing home values on gentrification likelihood may vary significantly depending on the concurrent changes in local income levels and educational attainment.

2. **Complex Feature Interactions**
   Random Forest's ability to model feature interactions was particularly valuable for our analysis. The model can automatically capture how different socioeconomic indicators work together to influence gentrification. For instance, the relationship between home price appreciation and gentrification might differ substantially across different metropolitan areas or income brackets—a complexity that Random Forest can naturally accommodate through its tree-based structure.

3. **Feature Importance and Interpretability**
   While logistic regression provides coefficients that can be difficult to interpret when features are correlated, Random Forest offers robust feature importance scores that help us understand which factors are most predictive of gentrification. This capability allows us to identify the relative influence of

Model Accuracy: 0.89

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.91	0.915	1039923
1	0.76	0.82	0.79	259995
accuracy			0.89	1299918
macro avg	0.84	0.87	0.85	1299918
weighted avg	0.89	0.89	0.89	1299918

FIG. 4: Model Performance

different socioeconomic indicators while accounting for their interdependencies.

4. **Feature Importance and Interpretability**
While logistic regression provides coefficients that can be difficult to interpret when features are correlated, Random Forest offers robust feature importance scores that help us understand which factors are most predictive of gentrification. This capability allows us to identify the relative influence of different socioeconomic indicators while accounting for their interdependencies.

5. **Robustness to Data Challenges**
Our dataset includes potential outliers, such as neighborhoods with extreme home price increases or unusual demographic shifts. Random Forest's ensemble nature makes it inherently robust to such outliers, whereas logistic regression would be more sensitive to these extreme values and potentially produce biased results.

F. Data Partitioning and Training

80% 20% test train split

sk learn library for RFE and the model

III. Results

The model achieved an overall accuracy of 89%, indicating strong predictive capability across both gentrifying and non-gentrifying neighborhoods. This high accuracy suggests that our selected features and model architecture effectively capture the patterns associated with neighborhood transformation.

The model showed varying performance characteristics across the two classes:

Non-Gentrifying Areas (Class 0)
The model exhibited particularly strong performance in identifying non-gentrifying areas:

- Precision of 0.92 indicates that when the model predicts an area will not gentrify, it is correct 92% of the time
- Recall of 0.91 shows that the model successfully identifies 91% of all non-gentrifying areas
- F1-score of 0.915 demonstrates strong balanced performance for this class

Gentrifying Areas (Class 1)
The model showed good, though slightly lower, performance in identifying gentrifying areas:

- Precision of 0.76 indicates that when the model predicts gentrification, it is correct 76% of the time
- Recall of 0.82 shows that the model captures 82% of actual gentrification cases
- F1-score of 0.79 reflects the balance between precision and recall for gentrifying areas

Aggregate Metrics
The model's performance can also be understood through its aggregate metrics:

- Macro average scores (precision: 0.84, recall: 0.87, F1: 0.85) provide an unweighted view of performance across classes
- Weighted averages of 0.89 across all metrics reflect the model's strong overall performance while accounting for class imbalance

The support values (1,039,923 non-gentrifying vs. 259,995 gentrifying cases) indicate a class imbalance in our dataset, which is expected given the relative rarity of gentrification. Despite this imbalance, the model maintains strong predictive performance across both classes. This performance profile suggests that the model is slightly more conservative in predicting gentrification, preferring to minimize false positives at the cost of potentially missing some gentrifying areas. This conservative approach could be valuable for stakeholders who prefer to avoid false alarms when identifying potentially gentrifying neighborhoods.

Figure 5 shows the colinearity matrix of the five features used in the model. They show no strong colinearity which is beneficial for the model.

IV. Conclusion

By understanding where and when gentrification is likely to occur, we can work toward solutions that benefit both new investments and existing residents—perhaps

	DP02_0064PE	median_home_value	DP03_0062E	DP03_0009PE	DP03_0135PE
DP02_0064PE	1.000000	0.569639	0.653088	-0.273869	-0.260668
median_home_value	0.569639	1.000000	0.592661	-0.143856	-0.137040
DP03_0062E	0.653088	0.592661	1.000000	-0.331882	-0.378287
DP03_0009PE	-0.273869	-0.143856	-0.331882	1.000000	0.233912
DP03_0135PE	-0.260668	-0.137040	-0.378287	0.233912	1.000000

FIG. 5: Colinearity Matrix of the Features Used in Model

```
zcta_predictions.head(10)
```

	ZCTA	year	predicted_gentrification
0	90011	2023	1
1	60623	2024	1
2	75216	2025	1
3	19140	2025	1
4	30315	2026	1
5	48213	2026	1
6	53206	2027	1
7	70117	2027	1
8	64127	2028	1
9	95838	2028	1

FIG. 6: Predicted ZCTAs. They are all located near big cities which mayb indicate a limitation of the use of ZTCA

finding ways to improve neighborhoods without wholesale displacement, or developing strategies to help long-term residents capture some of the economic benefits of neighborhood change through homeownership programs, small business support, or community investment trusts. However, this dual utility raises important ethical questions: Is gentrification an inevitable part of urban evolution? If so, how can we harness its economic benefits while protecting existing communities?

The challenge isn't just to predict change, but to help

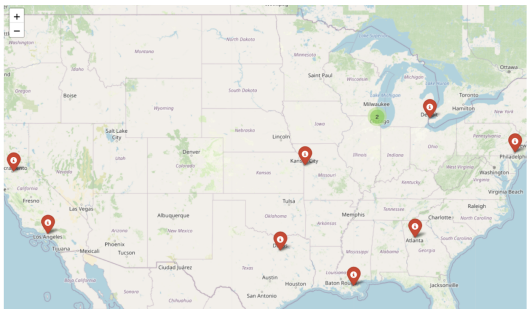

FIG. 7: Map of the predicted ZCTA areas

shape it. Whether gentrification is inevitable or not, our goal is to provide tools that can help create more equitable outcomes for all stakeholders involved in urban transformation.

[1] Bureau, US Census. "Relationship Files." Census.gov, 21 Nov. 2024, www.census.gov/geographies/reference-files/time-series/geo/relationship-files.2020.htmlzcta. Accessed 8 Feb. 2025.

[2] censusreporter. "Acs-Aggregate/Crosswalks/Zip$_t$o$_z$cta/ZIP$_Z$CTA *Aggregate.*" *GitHub*, 2020, *github.com/censusreporter/acs–aggregate/blob/master/crosswalks/zip$_t$o$_z$cta/ZIP$_Z$CTA$_R$EAD*