Cameron Sterling
Professor Wayne Lee
STATUN3106: Applied Machine Learning
March 14, 2025

Using Machine Learning to Understand the Prevalence of Rent Burdened Neighborhoods in NY Metropolitan Area

Problem

In the New York Metropolitan area, the cost of housing has become prohibitively expensive for a large segment of the population, leading to displacement, homelessness, financial strain, and preventing families from being able to save. This issue is especially acute for renters, who are often in lower income brackets with less financial stability.

The "cost burden" or "rent burden" is a critical metric to evaluate how expensive rent is by calculating monthly rent as the percentage of household income. Generally, a household is considered rent burdened if it spends more than 30% of its income on rent; a household that spends over 50% on rent is severely rent burdened.

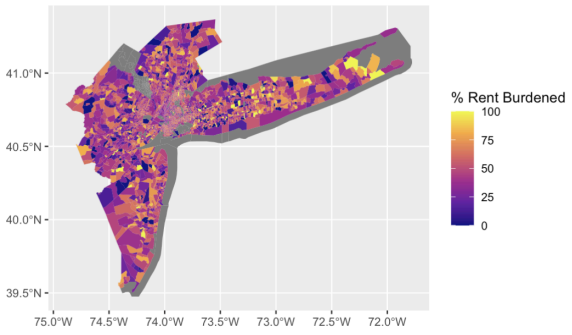


Figure One: Distribution of Rent Burden in NY Metro

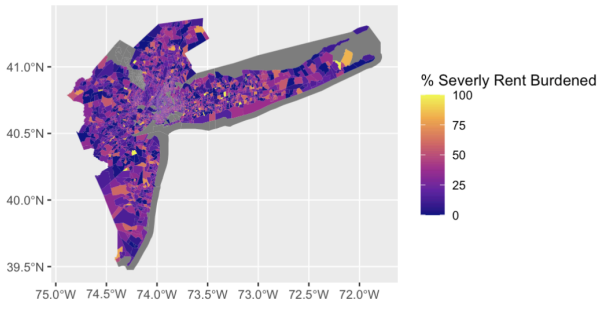


Figure Two: Severe Rent Burden in NY Metro

In a housing market context, we assume that individuals will maximize their utility – choosing housing options that will best fit their needs and financial means. However, when market forces

and structural barriers limit affordable choices, many households find themselves forced into unsustainable rent burdens. This project seeks to answer the question "why does rent burden vary across different neighborhoods in the New York Metropolitan area, and what demographic, environmental, and economic factors explain these differences?"

This question can be interesting for local policymakers and advocates seeking to understand why there are such extreme pockets of rent-burdened households in the NY-metropolitan area. It is important to ask if rent burden is happening because of cyclical poverty, if it is a reflection of households moving into higher-quality areas, or a mix of the two. One of the strongest results of this study was that the concentration of poverty in a neighborhood is an important predictor for rent burden, suggesting that a main contributor of rent burden is an overpriced housing market.

Data Sources

This project used three sources of data: the American Community Survey (ACS), Social Vulnerability Index (SVI), and Center for Disease Control's Environmental Justice Index (EJI).

American Community Survey (ACS)

The ACS is a nationwide survey conducted by the U.S. Census Bureau, providing detailed annual estimates on demographics, housing, income, and employment at various geographic levels. This project used a tract-level spatial resolution on 5-year estimates of the ACS from 2022. The ACS is the source of rent and housing tenure related variables.

Two key features were engineered from the ACS through using the percent of renting households sending a certain % of income on rent and dividing by total number of households:

- Percent of households rent burdened in a census tract (main outcome variable)
- Percent of households severely rent burdened in a census tract (main outcome variable)

Social Vulnerability Index (SVI)

The SVI is a composite index that measures a community's ability to prepare for, respond to, and recover from external stressors like natural disasters, economic disruptions, or pandemics. This project used the 2022 SVI on a census-tract level resolution. Instead of using the indexes, it uses

the features captured by the SVI, such as measures of socioeconomic status, household composition and disability, minority status and language, and housing and transportation related variables.

Environmental Justice Index (EJI)

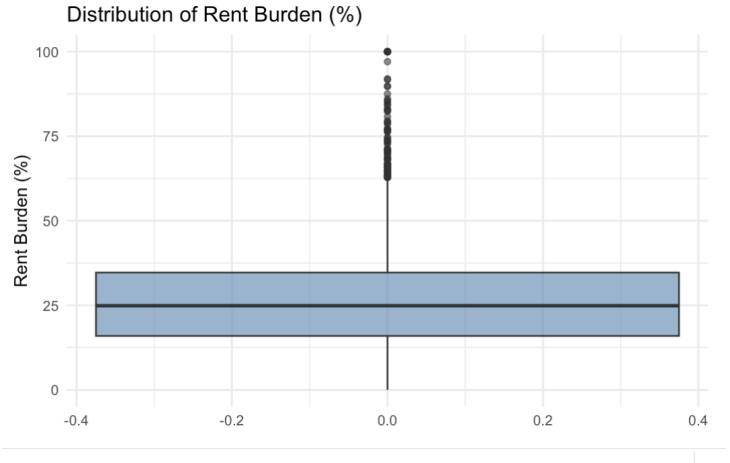
The EJI, developed by the CDC, measures cumulative environmental, social, and health burdens on communities, especially in disadvantaged areas on a census tract level. This project used some of the features related to environmental burden (air pollution, proximity to toxic waste sites) and health vulnerability (prevalence of chronic diseases) as features.

Data Quality

Since the goal of this project was to evaluate which features are the most important, we took a maximalist approach to adding features to the dataset. The final dataset included 4712 census tracts and 42 rows of data. After filtering the census tracts that had low data resolution (more than 30% of rows were NA), there were 4601 rows.

Since the resolution of the SVI did not include some NYC-specific resolutions of census tracts (NYC is so densely populated that some census tracts will be demarcated with decimals), we used a 2 KNN imputation to clear up NA's resulting from this issue.

The outcome variable of interest, Rent Burden %, had a wide distribution with the median percentage of households in a tract that was rent burdened at 43.47%. The distribution of % rent burden can be seen in the figure to the right. After initial data cleaning, there were no NA values.



Approach

This project considered two models: a linear model and a random forest. The linear model establishes a benchmark for understanding which factors might be correlated with rent burden, but it cannot be considered on its own because many of these features are correlated with each other. The presence of collinearity makes it difficult to determine the independent effect of each variable, limiting the model's interpretability and predictive accuracy.

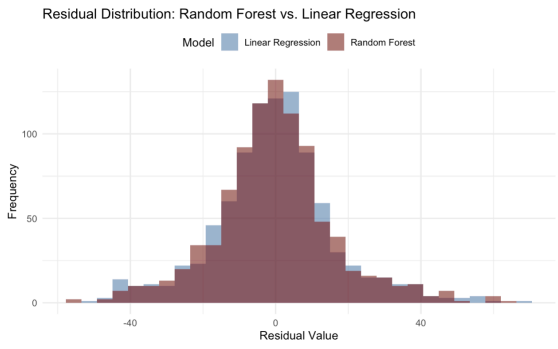
Theoretically, the random forest model should improve upon the linear model in feature selection while retaining its interpretability. Because random forest builds multiple decision trees using different subsets of the data and features, it can handle collinearity more effectively by distributing importance across correlated variables. Additionally, it can capture nonlinear relationships that a linear model would miss, potentially improving predictive accuracy.

To ensure that the results are robust and not overly dependent on a single train-test split, this project uses k-fold cross-validation with five folds on an 80-20 test split, which prevents overfitting. Through this approach, this project identifies some key drivers of rent burden in the New York Metropolitan area and assesses the relative importance of economic, social, and environmental factors.

Results

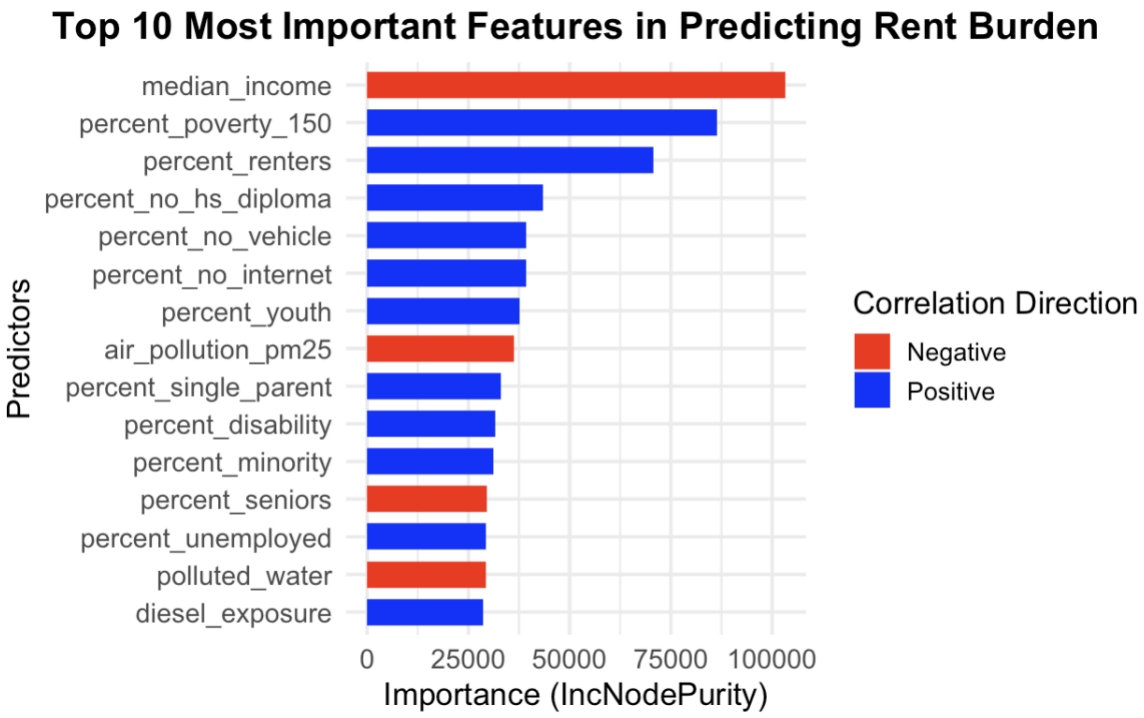
The cross-validated random forest model has an R^2 of 0.1811487, which outperformed the linear model with R^2 of 0.1507097. RMSE was similar between the models, at 16.51 from the random forest model and 16.808 for the linear model, suggesting a similar level of performance.

The figure to the right is a residual plot. Random Forest residuals (red) are more concentrated around zero compared to Linear Regression (blue), suggesting that Random Forest produces smaller errors overall. The Linear Regression model also has slightly more extreme residuals on both tails, indicating that it



struggles more with high-error predictions. However, we can also see that there are still many outliers which need to be explored.

Despite the relatively low R^2, the feature importance analysis revealed strong predictors of rent burden. The most important variables are shown in the figure below, as well as their linear correlation with the percent of rent burdened households in a neighborhood.



A cursory analysis of the results shows two interesting narratives. Firstly, the importance of median income in a neighborhood, percentage of people living in poverty, youth, and other demographic variables that usually indicate social vulnerability were important features in the model. This suggests that rent burden is concentrated in lower-income areas, where financial constraints leave households with fewer affordable housing options. High housing costs disproportionately impact low-income communities and reinforce cyclical affordability challenges.

The second, more unexpected finding, is the significance of some environmental factors, specifically air pollution (PM_25) and polluted water, in predicting rent burden. While random

forest does not inherently reveal causation, external correlation analysis shows a negative relationship between $PM_{2.5}$ exposure and rent burden. This might suggest that households are willing to take on higher rent burdens in exchange for improved environmental conditions, choosing neighborhoods with better air and water quality despite financial strain. This could indicate that environmental quality plays a role in perceived livability and housing demand, an area that warrants further investigation.

Limitations

While this project provides insights into the drivers of rent burden in the New York Metropolitan area, several limitations should be considered.

1. While Random Forest improves over linear regression in handling collinearity and nonlinear relationships, it has limited interpretability. Although feature importance rankings help identify key drivers, the exact nature of interactions between variables is not directly observable.
2. The data sources used—ACS, SVI, and EJI—are aggregated at the census tract level, which may obscure within-tract variability in rent burden.
3. Spatial autocorrelation could influence results, as rent burden is not randomly distributed but clustered geographically. Future research could incorporate spatial regression models or geospatial clustering to better capture neighborhood effects. The relative low R^2 of the final model indicates that there are other important factors to explore.

Conclusion