Ethan Carney and Abe Seybert
ECNS 460
Final Product

Introduction

We chose to do option A in performing a predictive analysis on the predictive capabilities of zoning classifications on housing price sales over time, utilizing the city of Bridgetown, CT. Our research question is defined as follows:

Research Question

"Can zoning classifications and their characteristics be used to predict housing price sales in Bridgeport, CT?"

We decided this question was important as city planners face many important decisions regarding re-zoning and zoning allowances that may have unintended spillover consequences (i.e. price changes in properties within the zone). Much of the previous research on this topic involves using specific price equations (Mark et. al, 1986), or standard economic analysis (Quigley et. al, 2005). Our application attempts to bridge the gap between these analyses and the recent developments of machine learning to explore further insights into the relationship between zones and property prices. This analysis is important in understanding how zoning regulations can affect the value of properties, and to what extent. Machine learning models are an excellent tool for exploring the ways in which zones and their characteristics can influence property sale values. They can pick up on relationships that may not be clear when an explicit model is applied to the data, so this application looks to provide additional insight on these potential associations. The results are relevant to city planners who will have a better understanding on how classifying a zone as a specific type can affect the properties in the area.

AI Disclosure

We utilized AI to handle debugging and error checking in our code. We also used it to generate ideas of certain plots to create and analyses to explore.

Our initial datasets came from two sources: A record of real estate sales in Connecticut that spans from 2001-2022 (from the official CT.gov website), and a GIS zoning data set with zone classifications and locations of zones in Bridgeport, CT (from the City of Bridgeport website). The reason for selecting Bridgeport is explained below.

Data Cleaning

Our initial datasets came from two sources: A record of real estate sales in Connecticut that spans from 2001-2022 (from the official CT.gov website), and a GIS zoning data set with zone classifications and locations of zones in Bridgeport, CT (from the City of Bridgeport website). The reason for selecting Bridgeport is explained below.

In the data cleaning process, we began by loading the housing sales dataset that contained records of housing sales in Connecticut from 2001-2022. Given the large size of this dataset (1 million+ observations), it seemed best to focus on one city, so we landed on Bridgeport, as it had the highest number of observations of any town. We then imported the GIS zoning data and visualized them using the viewer in R to better understand the spatial distribution of zones. The property sales data set had the following variables: house serial number, listing year, date the sale was recorded, town, address, assessed value, sale amount, sale ratio (Ratio of the property sales price to its assessed value), property type, residential type, non use code, assessor remarks, OPM remarks, and location (longitude and latitude points). There were a large portion of observations that had addresses for the property, but no location point. To address missing geo-spatial locations, we used the Google Geo-Spatial API to geocode addresses that didn't already have coordinates, implementing error handling for cases with incomplete or ambiguous addresses. By using parallel processing, we were able to handle the geocoding efficiently across a large number of records (although it was still computationally expensive).

The zoning data had the following variables: Name of zone (city designated code), Zone classification, and geometry (multipolygon of the zone's geographic location and shape).
This data set needed little cleaning, although we removed the altitude portion of the geometry to work with simply longitude and latitude values.

We then performed a spatial join of each property sale with the zone it was located in, and preserved both the property location point and the geometry of the zone it is located in. After combining the property data with the zoninging data, we refined the dataset further by removing unnecessary columns, such as "comments" and temporary variables used for geocoding. We organized the data so that each observation represented a unique property sale, and created a primary key based on the property's serial number and sale date. For consistency, we converted categorical variables to factors and truncated descriptions in the `Non.Use.Code` field, removing any leading zeros where necessary. We also set missing values to `NA`, replacing zero values in the assessed and sale amounts with `NA` to indicate cases where no assessment or sale occurred. Finally, we saved the cleaned dataset for further analysis.

The next data processing steps we took occurred after the "data visualization" phase in the project where we decided to pivot on our project topic slightly. Since we decided to take a machine learning approach to conduct a predictive analysis, we wanted to incorporate more variables (both zone related and geographic location related) to serve as potential predictors. We utilized geo-spatial data from the City of Bridgeport website that contained location data on important facilities/buildings in the city (city facility, university, community center, federal office court, medical facility, library, police station, etc.). We used this data to construct variables that quantify a house or property's proximity to important infrastructure, which appeared useful as easy access to these facilities is typically preferred. We only needed the facility type, the location, and the unique key (Parcel_ID) for our purposes. This data can be found in the "Neighborhood Assets.shp" file in the Raw Data folder of our GitHub. To process the data, we calculated the distance between each facility and each property we had sale information for. We then found the minimum distance from each property to each type of facility (since there were multiple facilities of the same type), and created new variables to store these distance measures. We then merged these variables with our already cleaned and processed property/zoning data.

The final data source we explored was manually constructed from the City of Bridgeport website. The website contained data for the types of structures/buildings that were permitted to be built within each type of zone. It was reasonable to include this data as the types of buildings located within a zone are likely indicative of the general price level of properties sold in that zone. We had to manually create a ".csv" file that contained binomial variables for each building type, indicating if it was allowed to be built in each type of zone (value of 1 if it was allowed, 0 otherwise). The building types were specified as: storefront building, commercial center, commercial house, general building, etc. We were able to directly merge this data set with the rest of the data by merging on the zone classification. We then converted the variables to factors with 2 levels. It should be noted that on the city website, there were a couple zone types that had no listed permitted buildings, as these zones had been reclassified. Thus, we gave these zones a value of 0 (or a level of 0), indicating that there were no permitted buildings allowed there.

Analysis

In our analysis of the variables, it became clear that both the assessed values and sale amount values for the properties were heavily right skewed. This made sense as there are some properties that have a much higher value than others, which skews the distributions. Thus, for our analysis, we utilized the log transformation of these variables. We then visualized the data to get a better understanding of how the property sales data and zoning data related. There was a notable discrepancy in the amount of sales data across zones, which was to be expected. Figure 1 highlights these differences.

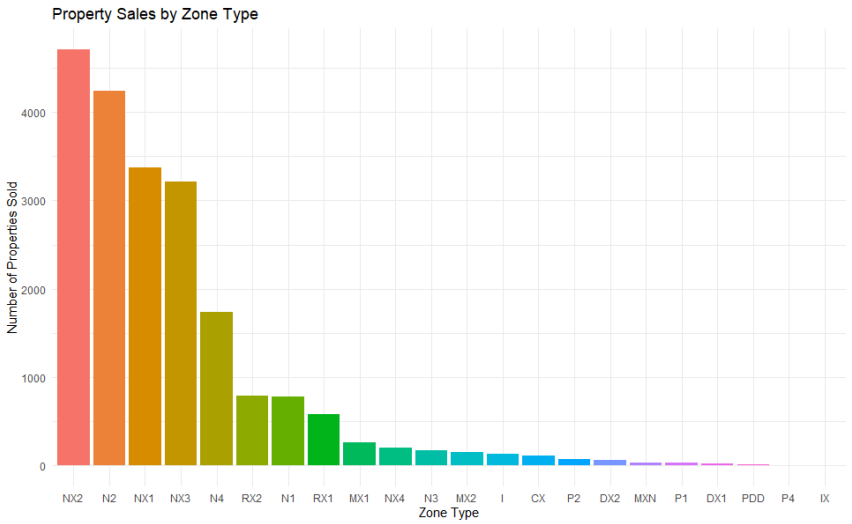


Figure 1. Number of Property Sales Across Zone Types

There is a much higher number of sales in the zones that begin with "N: and "R", which makes sense given that these are "Residential" and "Neighborhood" zones which contain more housing properties. This difference in observations across zones indicates that our analysis is better suited for domiciles rather than commercial use buildings.

We then looked to analyze the average property sale price across all zones in Bridgeport. We calculated the mean of the log of the sale price of properties sold in each individual zone, and the result is shown in figure 2 below.

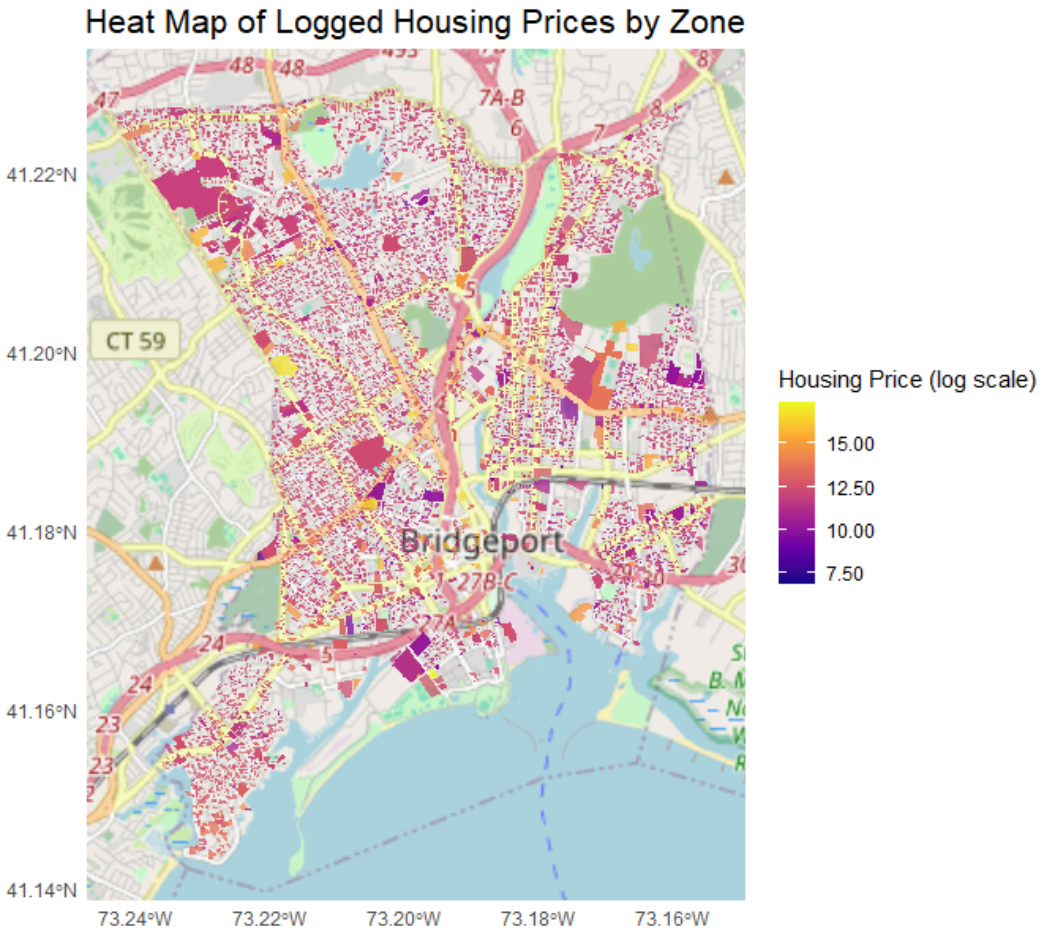


Figure 2.

Given the large number of zones and the size of the city, it is hard to visualize the historical sales data across zones. However, the plot does provide useful insights in that it shows a relatively uniform distribution of logged sale prices across zones in the city.

We then looked to perform a temporal analysis on the mean property sale price (log scale) across zone types over time. We also plotted the mean overall property sale price over time, as shown by the black line in figure 3.

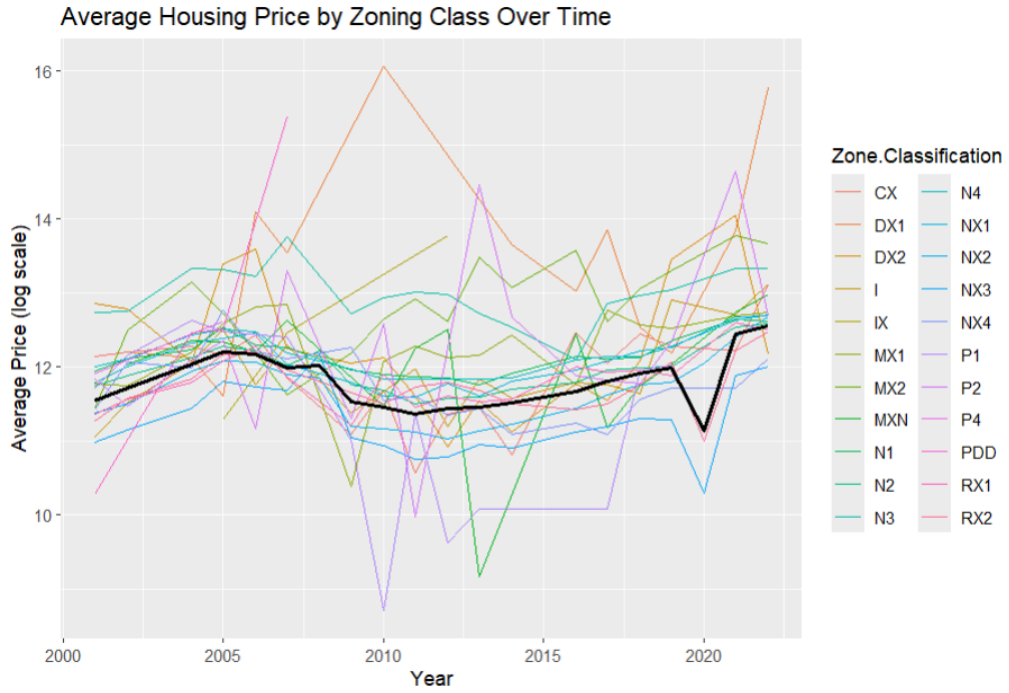


Figure 3.

The general shape of the mean logged sales price shows an increasing trend until about the 2008 crisis, after which it decreases until about 2010. After this point, it starts to slowly increase again, until about 2020, where it takes a sharp drop. This is indicative of the 2020 recession, and it was mainly driven by the drop in sale prices in the NX3 and CX zones. There is much more variability in sale prices across zones through time following the 2008 crisis. Overall, the changes in housing prices across time lines up with what we would expect, indicating the data has no notable unknown discrepancies.

We then checked if the property sale data was spatially correlated using a Moran test. We started by calculating the list of "neighbors" for each property, which was defined as any other property that is within a 0.01 km radius. We then created a spatial weight matrix, where each property was given a weight depending on its proximity to other properties. Using the "spdep" package, we then conducted the Moran's I test. The test checks whether the sale prices of properties exhibit spatial autocorrelation, or whether values of sale prices appear to be clustered in certain locations. The result of the test is summarized in the figure below:

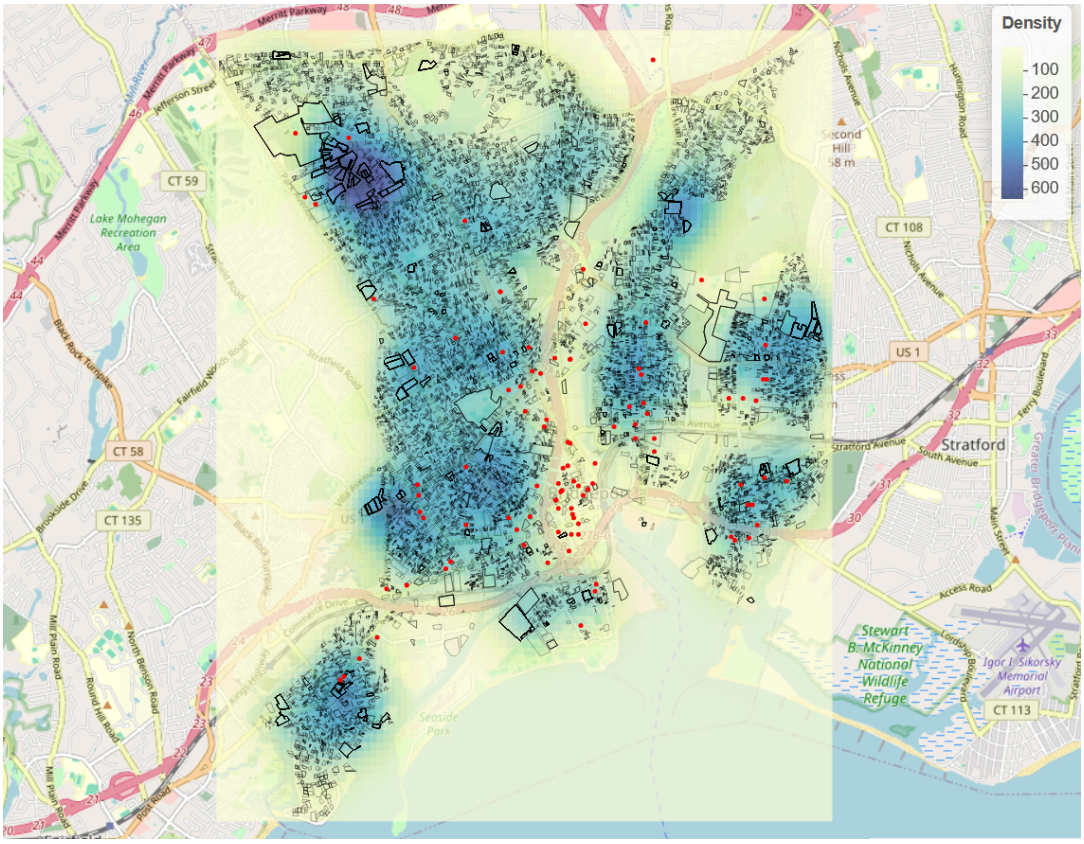
```
Moran I statistic standard deviate = 289.1, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     7.614103e-02     -4.834421e-05      6.945400e-08
```

Figure 4. Summary Results of Moran I Test for Spatial Autocorrelation of Property Sale Prices.

The test statistic was much larger than the expectation value, which is the expected value of the statistic under the null hypothesis where there is no spatial autocorrelation. Thus, the data showed significant spatial autocorrelation. The test also gave a very small p-value, indicating the presence of this correlation.

Our final pre-analysis before fitting our machine learning model involved using kernel density smoothing to visualize the density of sales across the city of Bridgeport. We utilized the "MASS" package to conduct a two-dimensional kernel density estimate over all property data. This was done using the "kde2d" function which took in the longitude and latitude values of each property and smoothed the points into a continuous surface. We then converted the surface to a raster object and used "leaflet" to visualize it in the RStudio viewer. We also plotted the facility locations (shown in red dots in the figure below) over the density smoothed surface to analyze any relationships between the volume of housing sales with proximity to important infrastructure buildings. By visual inspection, there appears to be a potential relationship between the density of the housing sales and the locations of the facilities. An output of the viewer is shown below in figure 5.

Figure 5. Kernel Density Smoothed Property Sales with Infrastructure Building Locations



Prediction

Our next task involved utilizing machine learning models to answer our research question. We decided to employ a random forest model, given that we had a large dataset with different types of predictors. We also wanted to avoid conducting an in depth feature engineering process, as there wasn't a clear indication from our analysis of how some of the variables would interact. To isolate the predictive capabilities of the zoning variables, we fit two random forest models: one in which the zone related variables were included in fitting and predicting, and one in which they were excluded. The "zone" variables included the factor variables of the zone classification and the building types allowed within each zone.

Results

The analysis we conducted offered insights into the predictive power of zoning classifications and their associated characteristics in determining property sale prices in Bridgeport, CT. The results from the random forest models, along with descriptive and spatial analyses, are summarized below.

Two random forest models were trained to predict (logged) property sale prices. We evaluated the models performance using the Root Mean Squared Error (RMSE) on the testing datasets as well as the training datasets, as shown in Table 1.

Model	Dataset	RMSE
RF with zone data	Train	300,146.23
RF with zone data	Test	491,843.91
RF no zone data	Train	311,921.44
RF no zone data	Test	721,549.18

Table 1. Zone data results and the RMSE of training sets and test sets.

The results show that the inclusion of zoning variables significantly improved the predictive performance of the models. The RMSE for the testing set is approximately 32% lower for the model including the zoning data ($491,843.91) compared to the model excluding it ($721,549.18). These results suggest that the zoning characteristics provide valuable information in predicting housing prices. Additionally, the performance difference between the two models was more pronounced on the testing set than it was on the training set, highlighting the importance of zoning data in generalizing predictions beyond the training set.

Despite the observed improvements, the RMSE for both the test data set and the training dataset had large RMSE values. These large errors can likely be attributed to the presence of large outliers in the housing dataset we used.

We attempted to address the issue of outliers by applying a logarithmic transformation to the sale price, but these values' effect was still notable after the transformation. In a dataset on house prices it is expected to have large values representing the unusually high-value properties in real estate data. By incorporating other variables like the zoning characteristics and additional predictors, our analysis reduced the impact of outliers. Future work could explore alternate methods for handling outliers, such as more robust regression techniques or manually trimming the extreme observations.

To assess the relative importance of zoning variables to the model that included them, we utilized the impurity-based importance. Features that result in a larger reduction in impurity when used in splits are considered more important, and a variable's overall importance is calculated as the sum of its impurity reductions over all trees in the forest. Figure 6 lists these importances.

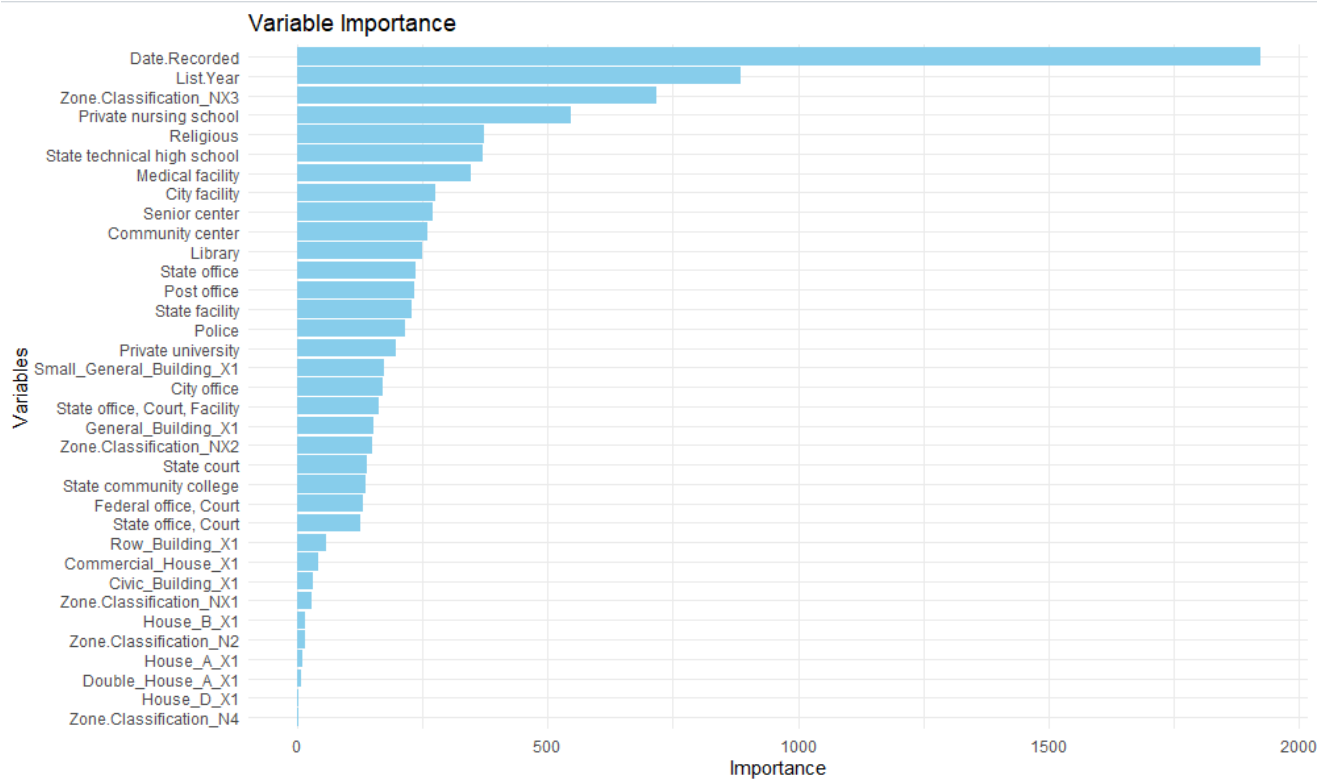


Figure 6. Variable Importance for RF Model Including Zone Features

It was clear from this analysis that the most important predictors were the temporal ones. However, the NX3 zone turned out to be the next most significant. Many of the remaining predictors of higher importance were the facility locations, but some of the allowed building/structure variables were closer to the "middle of the pack" with respect to their relative importance. It is clear that the zoning variables provided important predictive capabilities to the model.

Conclusion

This study examined the impact of zoning classifications on the property sale of housing units in Bridgeport, CT., using spatial analysis, descriptive analysis, and predictive modeling techniques. The primary goal of our analysis was to assess whether zoning and related attributes add value in pricing estimating models. Through data cleaning, feature engineering and spatial integration of datasets, we were able to build models to analyze the different variables. The results of the random forest models confirmed that zoning data is a useful predictor of property value. We can conclude that models incorporating zoning variables outperform those that do not have the zoning variables. Including the zoning variables reduces the RMSE by 32% on the testing set. Moreover, these findings show that zoning characteristics capture both spatial patterns as well as provide contextual information about property value.

In addition to analyzing zoning classifications, the analysis of proximity measures to infrastructure suggests that access to state run public amenities likely increases property value. While we did not explicitly isolate these factors in our models, the inclusion of them in the feature set demonstrates the opportunity for further analysis of neighborhood characteristics.

Overall, this study shows strong support that zoning data, when combined with spatial and neighborhood-level features, can provide significant and valuable insights in both understanding and predicting housing prices. Our results have many practical applications that can include informing policymakers, real estate agents, and investors.

Limitations

Since our analysis was confined to the city of Bridgeport, CT, we cannot conclude that zoning characteristics are necessarily predictive of property values in other cities throughout the U.S.. Bridgeport has had a unique growth when compared to the overall growth of the U.S., as there was a rapid growth from the mid 1800s to the early 1900s that has tapered off since then (CT, 2023). The majority of our property sale data was located within neighborhood and residential zones, so our results have stronger implications in these types of zones. Also, the NX3 zone classification had a strong importance in the model, but many of the other classifications had a smaller importance, or weren't selected as features at all. The ones that were selected were also mostly neighborhood zones, which seems to indicate that properties in these types of zones are better suited in using the model to predict their sale price. Since these zone classifications are unique to Connecticut, it is not reasonable to generalize the results to other regions that have a completely different set of zones and zone regulations. However, the results of the model indicate an important relationship between zones and property sale values that potentially exist in other regions as well.

References

Jonathan H. Mark, Michael A. Goldberg, *A study of the impacts of zoning on housing values over time*, Journal of Urban Economics, Volume 20, Issue 3, 1986, Pages 257-273.

Quigley, J. M., & Rosenthal, L. A. (2005). *The effects of land use regulation on the price of housing*. huduser.gov. https://www.huduser.gov/periodicals/cityscpe/vol8num1/ch3.pdf

CT. (2023, August 3). *Over time: Bridgeport's historical population - connecticut history: A cthumanities project*. Connecticut History | a CTHumanities Project - Stories about the people, traditions, innovations, and events that make up Connecticut's rich history. https://connecticuthistory.org/over-time-bridgeports-historical-population/

Zone Bridgeport. (2024). *City of Bridgeport zoning regulations.* Retrieved from

https://zonebridgeport.com/

Management, O. of P. and. (2024, September 4). *Real estate sales 2001-2022 GL: Connecticut Data*. CT Open Data Portal. https://data.ct.gov/Housing-and-Development/Real-Estate-Sales-2001-2022-GL/5mzw-sjtu/about_data