Analysis of 2011 Housing Market Topology Data for the City of Baltimore

Robert Brasso

***Abstract*— Home vacancies remain a major problem that many medium-large cities deal with. They use up valuable land, are often uninhabitable, cause a tax burden on the city, and bring down property values for the neighborhoods they are in. This paper explores the assumption that there is a relationship between home vacancy and the level of distress within their neighborhood. We will use a 2011 Housing Market Typology report from Baltimore Open Data to analyze whether or not certain classification algorithms can be implemented to classify the typology of a neighborhood with the hopes of being able to apply this algorithm in future studies to help predict neighborhood statuses based on real-time housing metrics.**

I. Introduction

In 1950, Baltimore's population was recorded at just under a million people and it ranked as the sixth largest city in the United States. Today, Baltimore's population is around 620,000 representing a 35% decline in the city's population over the last 60 years. The result of this dramatic decrease in population is that the city contains an inflated amount of vacant homes. According to an article in *The Atlantic* from 2014, Baltimore has upwards of 16,000 vacant homes and vacant homes outnumber homeless people 6 to 1. While the city has tried for years to address this issue, most policies have fallen short of making a large-scale difference in the problem. One philosophy to solving the problem of vacant homes is to increase the property value of the surrounding area, thus incentivizing people to move in and either renovate or replace these homes.

Using the 2011 Housing Market Topology dataset from Baltimore Open Data, we are attempting to create a model that accurately predicts the market category that a neighborhood will fall into based on a variety of other indicators.

In the paper that follows, we will provide a description of the dataset, results from our EDA (exploratory data analysis), develop a training and testing dataset, and then run three classification algorithms against the testing dataset to determine the best possible classification model for our experiment.

II. Dataset Description

The dataset used in this paper was taken from *Open Baltimore***,** which is a website operated by the city of Baltimore that provides free datasets relating to the city. The dataset is called "2011 Housing Market Typology" and was created in order to inform neighborhood planning efforts, also informing residents of the local housing market conditions in their communities. The entire dataset has 626 records, contains 12 fields, and can be found by going to the following link: https://data.baltimorecity.gov/Housing-Development/2011-Housing-Market-Typology/782b-zpd7. Below each of the fields will be described in detail.

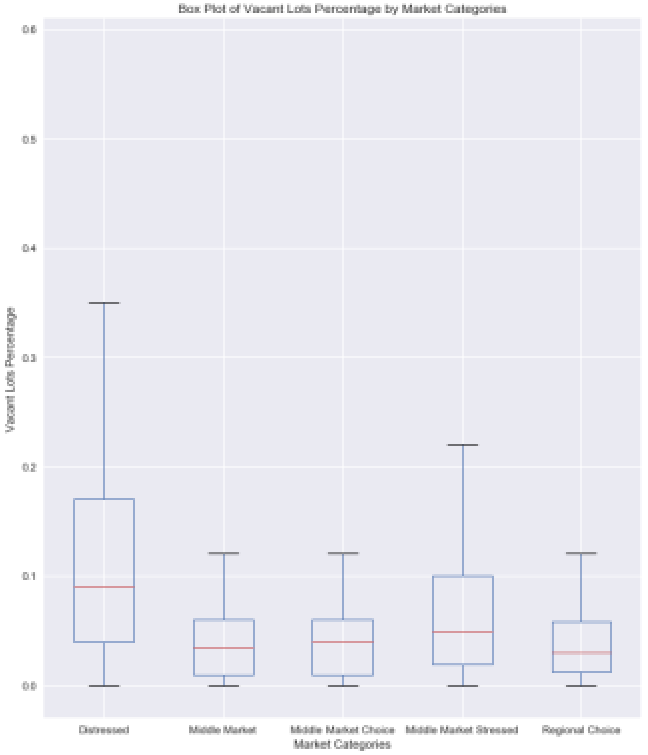
- **blockGroup** – Census block group. A map of all Baltimore census block groups can be found at http://www.mdp.state.md.us/msdc/census/cen2010/maps/blkgrp10/Baci_blkgrp10Roads.pdf
- **marketCategory** – Each block group is listed as one of the following groups…
    - **Regional Choice** – Competitive housing markets with high owner-occupancy and high property values
    - **Middle Market Choice –** Housing prices above the city average with strong ownership rates, low vacancies, but slightly increased foreclosure rates
    - **Middle Market –** Median sales of $91,000 as well as high ownership rates. Higher foreclosure rates, with slight population loss.
    - **Middle Market Stressed –** Slightly lower home sales than city average and have not shown significant sales price appreciation. Vacancy and foreclosure rates are high and the rate of population loss has increased.
    - **Distressed –** Experienced deterioration of housing stock. Contains high vacancy and the lowest homeownership rates. Most substantial population loss.
- **sales20092010 –** Total number of residential sales from 2009 to 2010.
- **salesPriceCoefficientVariance –** Sales price standard deviation from 2009-2010 divided the mean sales price from 2009-2010.
- **commericalResidentialLandRatio –** Commerical and institutional land area divided by residential land area.
- **unitsPerSquareMile –** Amount of housing units per square mile.
- **residentialPermits –** Residential permits greater or equal to $50,000 divided by residential lots plus vacant lots.

- **vacantLots –** Amount of vacant lots divided by residential plus vacant lots.
- **vacantHouseNotices –** Vacant housing notices divided by residential plus vacant lots.
- **foreclosureFilings –** Foreclosure filings from 2009-2010 as a percentage of privately owned residential lots.
- **medianSalesPrice20092010 –** The median sales price for homes in census block from 2009-2010.
- **ownerOccupied –** Estimation of all occupied residential units that are owner occupied.

III. Exploratory Data Analysis

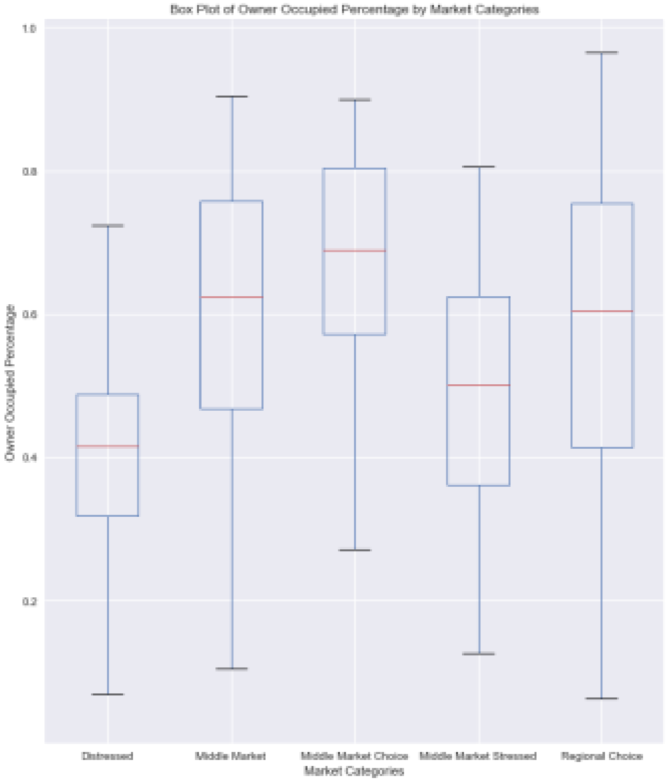
Beginning our exploration of the dataset, the first piece of information we needed was the breakdown of the amount of records for each of the market classifications. 209 of the census blocks were reported as Distressed, 117 as Middle Market Choice, 114 as Middle Market Stressed, 94 as Regional Choice, and 92 as Middle Market.

The next task was to take a look at all of the specific attributes and identify how they interact with each other. In figure 2.1 (end of paper) you will see a scatter matrix of all of the attributes. The attribute names have been adjusted slightly for readability. From a review of the very large scatter matrix, there were a handful of relationships I wanted to explore further. Below are a series of box plots used to analyze these relationships.

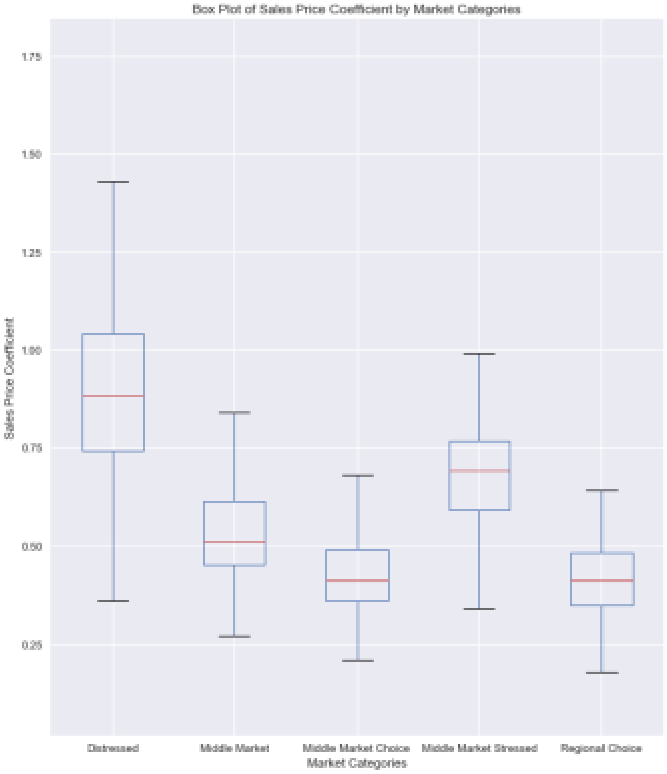


The first relationship examined was between market categories and Vacant Land Percentage. Based on the above figure, it appears there is a higher percentage of vacant homes in the distressed and middle market distressed neighborhood blocks.

Additionally, on the next boxplot we see that the owner-occupied rates for those neighborhoods is significantly less than the others.



Additionally, it appears that the sales price coefficient may be a significant indicator of distressed and middle market stressed communities.



While there are likely more relationships that could be identified, it is fair to say that the attributes in this dataset provide some insight to the classification they have and therefore will be a good candidate for our classification algorithm testing.

IV. Classification Models

Our goal with testing the following classification algorithms is to isolate and algorithm that will accurately classify housing categories based on a variety of statistics. Once we have identified the best algorithm, we can identify how the values entered into the algorithm impact the classification, however that is beyond the scope of this paper. Future research can use this algorithm to identify the key indicators that describe neighbors and attempt to provide early detection when neighbors begin to fall into distressed and middle market distressed categories

For testing this data against the classifier algorithms, we used a 70/30 split to create our training and testing datasets. The three algorithms we will be analyzing are K Nearest Neighbors, Random Forest, and Naïve Bayes. Below are descriptions of the results from each of our algorithm testing

A. K Nearest Neighbors

For K Nearest Neighbors we set the distance measure as Euclidian distance and first set N = 2. The initial results from N = 2 were promising, but in order to maximize the accuracy, recall, and precision, we wanted to test how various N values impacted the results. Below is a chart of the N values and their corresponding metrics.

N - Value	Accuracy	Precision	Recall	F-Score
1	.9521	.95	.95	.95
2	.9521	.95	.95	.95
3	.9468	.95	.95	.95
4	**.9628**	**.96**	**.96**	**.96**
5	.9521	.95	.95	.95
6	.9521	.95	.95	.95
7	.9415	.94	.94	.94

After testing several values of N, we see that the best results came from N=4. While the algorithm clearly works well with all values of N under 7, performance appeared to be optimized at N=4. The lowest performance was when N=7 and leads me to believe that even higher values of N would result in lower scores. To test that, I re-ran the algorithm with N=50 and while the scores were lower than N=4, they were almost identical to N=7.

B. Random Forest

Similar to our testing for K Nearest Neighbors, we will make a chart of the accuracy, precision, recall, and F Scores where the N_Estimator value is changed. Below is our test results for Random Forest.

N_Estimator	Accuracy	Precision	Recall	F-Score
5	.8830	.89	.89	.89
10	.8883	.89	.89	.89
50	.9149	.92	.91	.91
100	**.9521**	**.95**	**.95**	**.95**
150	.9309	.93	.93	.93
200	.9309	.93	.93	.93
300	.9309	.93	.93	.93

For the Random Forest algorithm, it appears that the best we can do is .9521 accuracy with a .95 f-score. While these values are very strong, they fall short of K Nearest Neighbors when N=4. Our optimized value for Random Forest appears to be when N_estimator = 100.

C. Naïve Bayes

Plugging our training and testing datasets into the Gaussian Naïve Bayes model, we received the following scores; accuracy = .9309, precision = .94, recall = .93, and f-score = .93.

V. Conclusion

After reviewing all of our data it became evident that all of the algorithms we tested do a fair job at classifying data into the various housing categories. However, the clear winner of our testing is K Nearest Neighbors which got us up to a .96 f-score with .9628 accuracy. Beyond that, K Nearest Neighbors was also the most accurate in all of those scores for each of the individual classification categories. Comparatively, while the Naïve Bayes model was close in terms of overall scores, it performed relatively poorly when classifying regional choice, providing a precision score of .79, which was the lowest score seen for any category during this testing. This leads me to believe that the best algorithm of the ones we tested for this dataset is K Nearest Neighbors, while Naïve Bayes was the worst.

References

https://www.theatlantic.com/business/archive/2014/10/can-homeless-people-move-into-baltimores-abandoned-houses/381647/

Figures

2.1