K Nearest Neighbor Clustering to Identify Developing Neighborhoods in Baltimore, MD

Andrew Timmons, Coursera Capstone Project

Abstract

The identification of neighborhoods that are candidates for future growth and gentrification is central to effective long-term real estate investment. This is particularly true in American cities, many of which have received a large influx of population resulting from the 2008 financial crisis which destabilized many suburban and rural communities. The availability of public datasets and the analytical capacity of machine learning algorithms present a unique opportunity to quantitatively identify those neighborhoods which are poised for rapid development. Here, I present an example case in the city of Baltimore, Maryland. As a recent graduate of a PhD program at Johns Hopkins University, this topic is particularly prescient as I am eager to purchase a home in Baltimore City. In this report I will detail my use of machine learning to identify affordable neighborhoods that are poised for further development, resulting in healthy returns on my real estate investment.

Introduction

Few American cities exemplify the pervasive issue of 'urban blight' to the extent of Baltimore City, Maryland. Located in the notorious 'Rust Belt,' Baltimore has experienced substantial economic and population collapse in the past 50 years. Plagued by contractions in the American steel industry, since 1970 the population of Baltimore City has declined by >30% (Fig. 1, U.S. Census). Contaminant with this rapid economic and demographic decline, the crime rate in Baltimore City has skyrocketed, giving Baltimore the dubious distinction of one of the most violent cities in the United States.

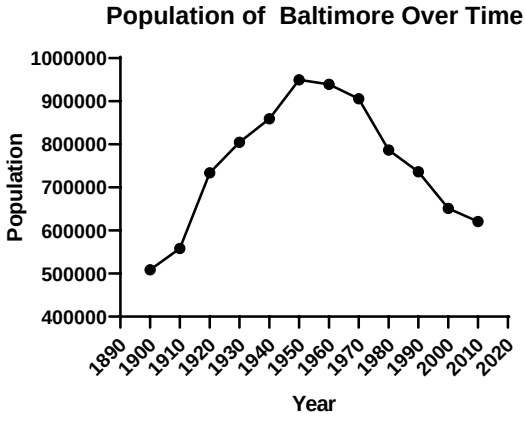

Fig. 1 – Population of Baltimore City from 1900-2020. Data taken from US Decennial Census

Vast financial and social resources have been deployed the revitalize some of Baltimore's most destitute neighborhoods. As a result of public and private investments, certain neighborhoods have begun experiencing a revitalization of neighborhood character and local business. Early investment in a gentrifying neighborhood is a potentially lucrative opportunity, particularly for young professionals purchasing a first home. Fortunately, with the variety of data available from the Baltimore City government, the US census, and the FourSquare venue database, we can identify candidates for gentrification using quantitative, machine-learning approaches.

Methods

Using K Nearest Neighbor (KNN) clustering approaches, I will identify neighborhoods with low median property values in Baltimore that show similarity to more expensive neighborhoods. Baltimore has 55 recognized neighborhoods, which represent a wide spectrum of home prices, crime rates, and neighborhood amenities. Clustering will consider a variety of different venues as taken from FourSquare

and various demographic statistics provided by the Baltimore City Government. The desired output from the KNN clustering approach is to identify less expensive neighborhoods that are similar in character to more expensive neighborhoods.

Because KNN clustering predicts discrete categories, I will coerce the target metric (the median home sale price in each neighborhood) into discrete categories by 'binning' according to the median home price. Using all 55 neighborhoods, I will calculate 5 evenly 'bins' according to median home price, and label each neighborhood accordingly (i.e. bin 0 will contain those neighborhoods in the 0^{th}-20^{th} percentile of median home prices). A KNN clustering instance will be trained on a subset of Baltimore neighborhoods, and the accuracy in predicting the categories of neighborhoods in a test set will be assessed. Finally, I will individually inspect each cluster to identify less expensive neighborhoods that display similarity to more expensive neighborhoods.

Throughout this work, I will use several datasets which I have detailed below

Data Set 1: Baltimore Demographics

The United States census provides neighborhood-level statistics on total populations across the United States. I will source this data from the United States census bureau for each neighborhood within Baltimore city. I will directly import this data into a dataframe from a local file to avoid navigating an API call to a 3rd party dataset. All data files used will be stored in my github repository.

Data Set 2: Baltimore Crime Rates

The Baltimore City police department provides a list of crime statistics. Using this dataset, I will quantify the number of violent crimes occurring in each neighborhood.

Data Set 3: Venues in Baltimore City

Using the FourSquare dataset, I will quantify the number of different venues occurring in each neighborhood. Using these venues, I will use KNN algorithms to cluster neighborhoods in Baltimore city.

Results

Initial Inspection of Data

Demographic data were imported from the Baltimore OpenData project, facilitated by ArcGis. Shown below is a selection the demographic data that was obtained for all 55 neighborhoods in Baltimore.

	Neighborhood	Population	Crimes per 1000 People 2018	Median Income in 2018	Median Home Price in 2018
0	Allendale/Irvington/S. Hilton	16217	20.657335	38535.562176	75000.0
1	Beechfield/Ten Hills/West Hills	12264	12.312459	58055.306613	159450.0
2	Belair-Edison	17416	15.215893	42633.619512	79900.0
3	Brooklyn/Curtis Bay/Hawkins Point	14243	24.362845	39936.512500	67500.0
4	Canton	8100	9.135802	116911.088235	295500.0

Fig. 2 – Example of demographic data format

Before obtaining venue data for each neighborhood, I wanted to assess the relationship between the demographic metrics that were imported for each neighborhood. As demographic metrics are often

associated, it will be important to know how demographic-to-demographic relationships will affect our clustering algorithm. As shown in Fig. 3, there is a high level of correlation between median income within a neighborhood and the median home price. This is expected, as more affluent citizens will live in more affluent neighborhoods. A secondary, negative correlation was seen between median incomes, median home prices, and the crime rate per 1000 neighborhood residents, highlighting that wealthier neighborhoods typically experience less crime than poorer neighborhoods. The correlation between crime rate and median home prices is important in a practical sense for this project, as I would like to avoid buying a home in a high-crime area. With this information, I will be most interested in those neighborhoods that correlate with expensive neighborhoods, as these neighborhoods will most likely have less crime than neighborhoods that cluster on the lower end of lower median home price categories.

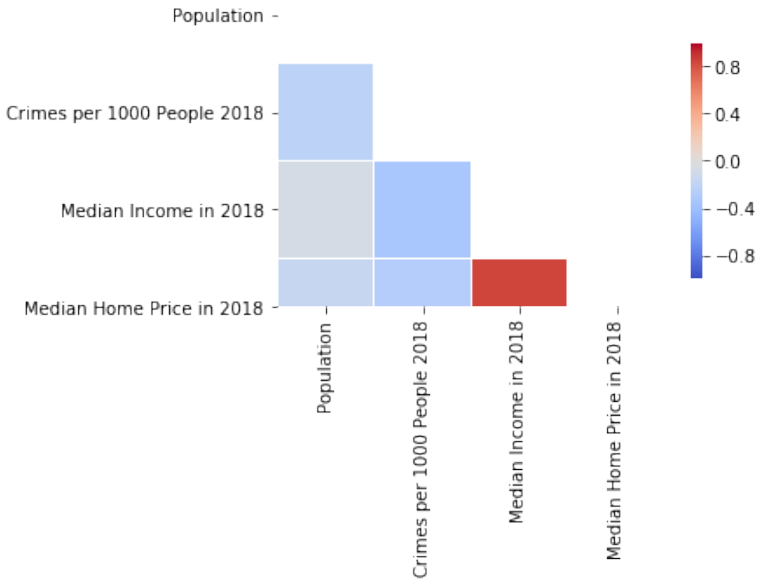

Fig. 3 – Clustering heatmap of demographic data

To obtain a better understanding of different neighborhoods in Baltimore, it is important to use more than demographic data obtained from the US Decennial census. Fortunately, FourSquare provides a rich repository of venue data within each neighborhood. To enhance this analysis, I used FourSquare to obtain the numbers of different cultural, practical, and educational amenities present within each neighborhood in Baltimore. By including this data in the clustering algorithm, I intend to get a more nuanced understanding of how different neighborhoods in Baltimore relate to each other.

To obtain venue data for each neighborhood, I need to prepare my data to be used in the FourSquare API. The initial required step is to convert neighborhood names to latitude and longitude coordinates. This was performed using reverse geocoding facilitated by the 'geopy' python module. Latitude and longitude coordinates were obtained for each neighborhood and appended to the datasets already collected (Fig. 4).

	Neighborhood	Population	Crimes per 1000 People 2018	Median Income in 2018	Median Home Price in 2018	Latitude	Longitude
0	Allendale/Irvington/S. Hilton	16217	20.657335	38535.562176	75000.0	39.2816626	-76.6840405
1	Beechfield/Ten Hills/West Hills	12264	12.312459	58055.306613	159450.0	39.2878195	-76.7003502
2	Belair-Edison	17416	15.215893	42633.619512	79900.0	39.3215607	-76.5678685
3	Brooklyn/Curtis Bay/Hawkins Point	14243	24.362845	39936.512500	67500.0	39.2257902	-76.5892145
4	Canton	8100	9.135802	116911.088235	295500.0	39.2816434	-76.5730592

Fig. 4 – Latitude and longitude coordinates for Neighborhoods in Baltimore

FourSquare maintains up-to-date data on hundreds of types of venues. Because I can submit a limited number of requests to the FourSquare API, I selected a subset of venue categories that I thought would be most informative. The categories chosen are liquor stores, grocery stores, schools, coffee shops, libraries, hotels, Italian restaurants, and bistros. Using a series of calls to the FourSquare API, I obtained the number of each of these venues within 800 meters of the center of each neighborhood. I chose 800 meters, as this is the average length of 10 American city blocks and represents a comfortable walking distance.

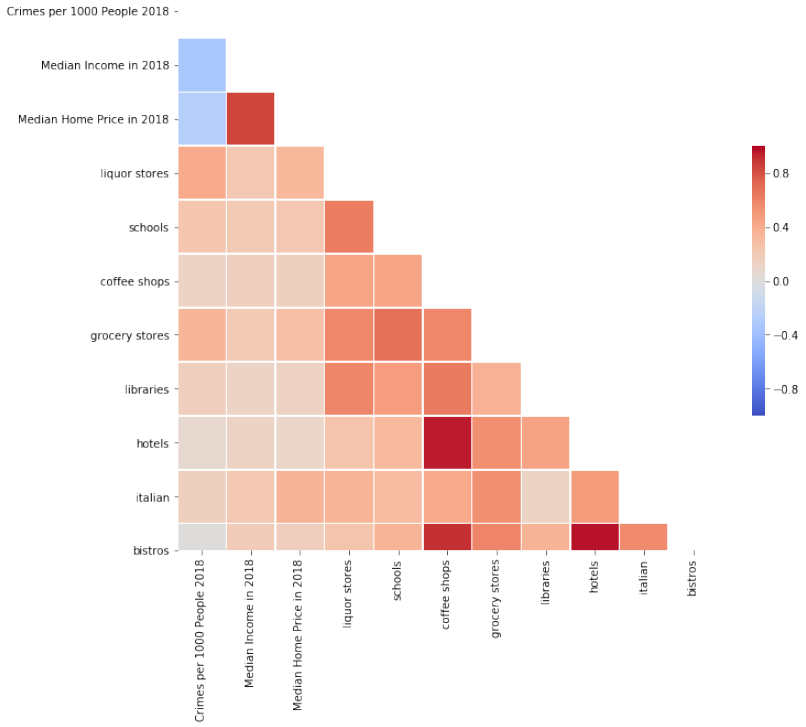

Fig. 5 – Correlations between venue distributions

As with demographic data, it is pertinent to understand correlations between venue distributions. I performed this analysis in a similar fashion as the demographic data to observe all correlates within the dataset (Fig. 5). Interestingly, pronounced correlations were observed between certain venues, particularly between hotels and coffee shops, which had a pearson correlation greater 0.8. Further, there is a slight positive correlation between many different types of venues, and the median home price. While none of these individual correlations is strong enough to be predictive, when used as inputs in a KNN clustering algorithm these inputs may help to enhance the accuracy clustering.

Now that all datasets are obtained, the clustering algorithm can be built and trained on a subset of the neighborhood data. Data were divided into independent (all metrics except median home price) and dependent datasets (median home price). To negate undue influence of any variable in the independent dataset, data were normalized and centered around 0. Following normalization, data were divided into testing sets and training sets. Our training set contained 45/55 neighborhoods in Baltimore, and our testing set contained 11/55

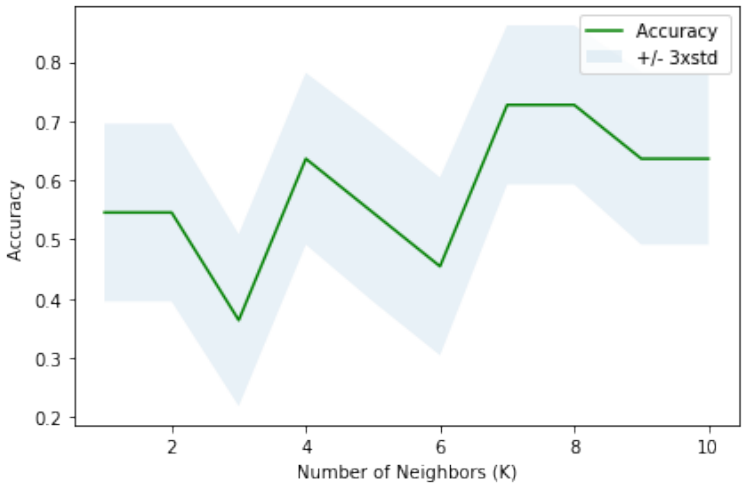

Fig. 6 – Accuracy scores of the KNN clustering algorithm with different K-neighbor parameters

neighborhoods. One neighborhood, the Waiverly neighborhood, was not recognized by the FourSquare API and was omitted from analysis.

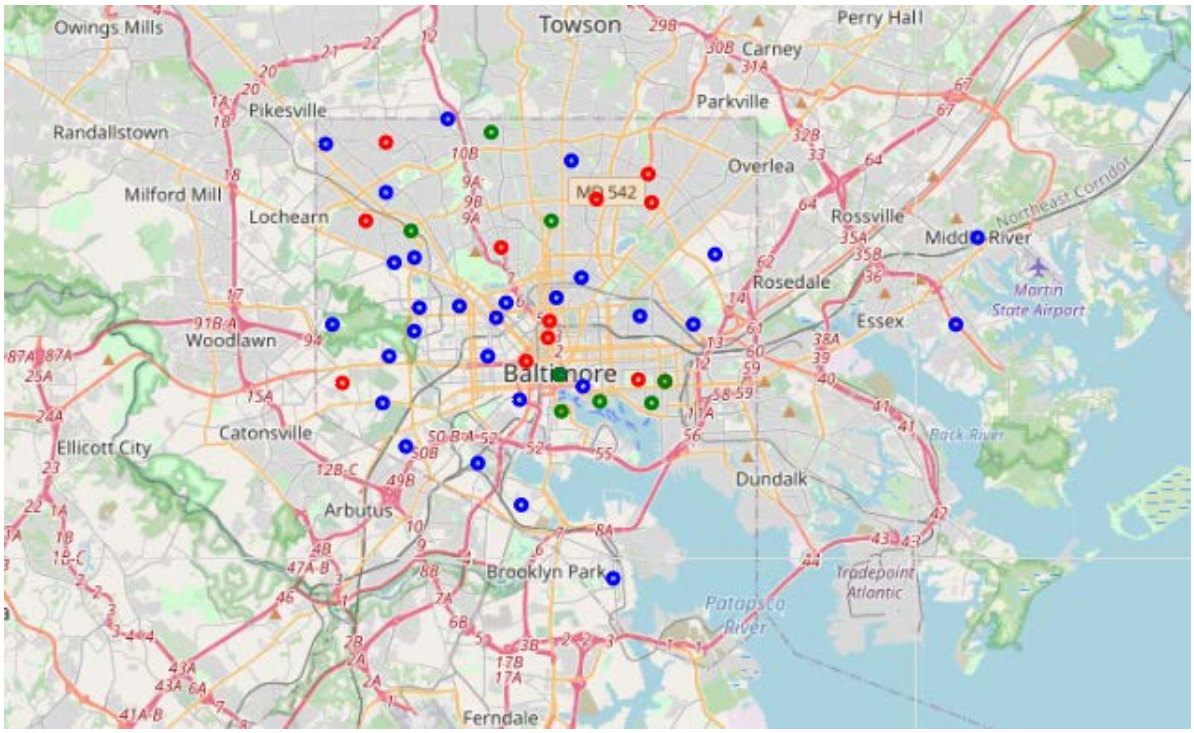

Fig. 7 – Distribution of neighborhood clusters throughout Baltimore

Given that our input datasets were relatively small, we were able to perform significant optimizations to our KNN clustering algorithm without expending large amounts of computing power. Using a series of iterative loops, we tested several different k-neighbor values to identify which resulted in the best clustering performance and accuracy. We identified that a k-neighbor value of 7 resulted in an overall accuracy score of 0.723 (Fig. 6). Using this trained KNN clustering instance, I clustered all 54 available neighborhoods in Baltimore.

The 54 analyzed neighborhoods were placed into 3 distinct clusters. When overlaid on a map of Baltimore, these neighborhoods appear evenly distributed throughout Baltimore (Fig. 7). However, cluster 1 (blue dots) has substantially more neighborhoods than either cluster 2 (red) or cluster 3 (green). This is most likely due to the relatively few numbers of neighborhoods in Baltimore that are currently experiencing development and gentrification.

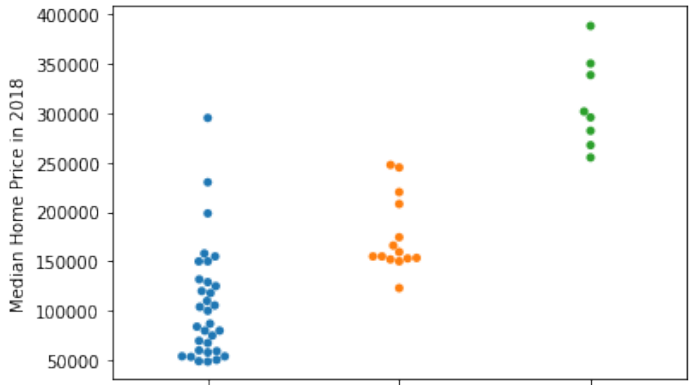

Fig. 8 – Median income prices within each cluster

Each cluster was individually inspected to identify the distribution of home values within each cluster. As expected, each

cluster represents generally different levels of home median home prices, with cluster 3 representing the highest home prices in the city. Fortunately, cluster 3 shows a relatively wide range of home values, which may enable the identification of cheaper neighborhoods that are similar in character to more expensive neighborhoods.

	Neighborhood	Population	Crimes per 1000 People 2018	Median Income in 2018	Median Home Price in 2018
37	Mount Washington/Coldspring	5168	5.224458	84848.090909	255000.0
15	Fells Point	9039	19.028654	98981.910064	267500.0
27	Highlandtown	7250	31.034483	88324.005277	282000.0
4	Canton	8100	9.135802	116911.088235	295500.0
29	Inner Harbor/Federal Hill	12855	14.080124	100932.230594	301500.0
38	North Baltimore/Guilford /Homeland	17464	4.466331	98094.852941	338350.0
47	South Baltimore	6406	4.995317	114107.319502	350000.0
21	Greater Roland Park/Poplar Hill	7377	2.575573	117951.409574	388000.0

Fig. 9 – Neighborhoods sorted by home value in cluster 3

To identify these neighborhoods, I selected only those neighborhoods included in cluster 3, and sorted the dataset by median home price Fig. 9. Inspection of this dataset reveals that the Mount Washington/Coldspring, Fells Point, and Highlandtown areas cluster with the more expensive neighborhoods of Roland Park and South Baltimore. Interestingly, the Mount Washington/Coldspring, Fells Point, and Highlandtown areas are, on average, less than 75% of the cost of the more expensive neighborhoods of Roland Park and South Baltimore. Visually inspecting this cluster reveals that, despite having substantially lower home prices, the Mount Washington, Fells Point, and Highlandtown areas share low crime rates and have similar venue distributions to the more expensive neighborhoods of South Baltimore and Roland Park.

Discussion

To identify neighborhoods that represent attractive investment opportunities, I wanted to identify neighborhoods in Baltimore that were like expensive neighborhoods, but with lower home values. In addition to enjoying an elevated quality of life, it is reasonable to expect that neighborhoods that resemble expensive neighborhoods may themselves experience an increase in property values.

I performed quantitative analysis using K Nearest Neighbor clustering to identify those neighborhoods that are similar in demographics and neighborhood amenities to expensive neighborhoods. Using the FourSquare repository, I obtained the quantity of many different types of venues in Baltimore city, including libraries, schools, coffee shops, etc. By including this venue data for each neighborhood, I was able to provide more nuanced data to the clustering algorithm than generalized demographic data.

Conclusion/Future Directions

From this analysis, I would assume that the Mount Washington, Fells Point, and Highlandtown areas are the most attractive neighborhoods in which to purchase a house Despite having substantially lower home prices, the Mount Washington, Fells Point, and Highlandtown areas share low crime rates and have similar venue distributions to the more expensive neighborhoods of South Baltimore and Roland Park.

This analysis can be refined by including more venue data from FourSquare and more demographic data from the Baltimore City opendata project. Additionally, repeating this analysis with different machine learning algorithms (such as logistic regression) will allow inspection of factor coefficients. This will permit inference on the relative impact of each factor on a neighborhood's home prices