Neighborhood Performance in Post-Recession Detroit

By Nathan Schumer

The city of Detroit is the site of competing narratives about its fate and meaning. First, to state the obvious, Detroit was an industrial capital of mid-20th century America, which sprawled and suburbanized, driven in large part by racial conflict and the 1967 Detroit riot, as well as subsequent conflicts over school busing, which provided deep racial undertones to the relationship between the city and its surrounding suburbs. As the auto producing capital of the country, Detroit dramatically removed its previously existing transit system, opting for ever outward expansion with automobiles for all. More so than any other city, Detroit also embodied the promise of 1950s America, that a high school degree could lead to a high paying union job with a pension, vacation, and enough money for a single earner to support a household. All of this began to collapse in the 1970s, as changing industrial patterns drove down automaker profits, and conservative political regimes began to undermine union power. Austerity drove down investments in the public sector, and racial segregation (along with a spiking crime rate) led to a vision of Detroit as an urban wasteland. The late 1980s and 1990s began to see the steady erosion of the earlier industrial regime, and the 2000s saw its near collapse during the financial crisis. Detroit contracted throughout this time period, as its population steadily bled away. The loss of population led to a loss of revenue for the city, which was saddled with the obligations of an earlier era, and Detroit was forced to consider how it would respond to its residents and its pensioners, as the cycle of population loss and large pension obligations continued.

The city is 142.9 square miles, with a population in 2016 of around 670,000. The population density of the city has dropped with each census from a high of 13,330 per square mile in 1950 to 5,144 per square mile in 2010. The city's neighborhoods have emptied out, as blighted areas and abandoned houses have slowly taken over. The city is also faced with the impossible task of offering services to vast areas with few residents or trying to find ways to concentrate residents into existing areas to provide services. Furthermore, the eroding tax base and limited services of the city made the downtown of the city relatively unattractive to most businesses. Yet, in the 2010s and following, business concentration in the city has grown, as a new cohort of younger people have increasingly chosen to live downtown.

There are two main ways of describing the fate of Detroit. Some narratives describe dramatic revitalized city, which has turned a postindustrial landscape into a new city, bouncing back from the recession and the bankruptcy of the city, with a leaner, more flexible government. This narrative portrays Detroit as triumphant in the face of decades of neglect and mismanagement, a renaissance fueled by strong growth in Detroit neighborhoods, as the city undergoes a process of downsizing and rightsizing.[1] Other narratives of the city tell a story of neglect, displacement, and gentrification, where longtime residents have been pushed out or swept under the rug to make way for new neighborhoods with higher tax dollars. Detroit is so large and vast that in fact, both stories can be true, and one aim of this project is to describe and predict some of the different facets of performance between different neighborhoods, identifying how and why some neighborhoods succeeded and others did not.

In general, this project is concerned with the distinctions between rich and poor neighborhoods, using the specific metrics of mean income, total buildings permitted, crime and homicides, auctions (functionally, proxies for displacement), and Improve Detroit tickets (a database for reported issues in a neighborhood) and their wait times. The specific goal is to see if there is strong evidence of this wealthy/poor neighborhood divide, and to look at the specific factors that might contribute to such a division. Are

[1] See, for instance, https://www.usatoday.com/story/opinion/2015/07/06/fixing-detroit-mismanagement-housing-transit-revitalization-strategy-struggling-us-cities-column/29577383/

differences in equity a function of material conditions? Does the city of Detroit administer all areas equally? What are some proximate causes of growth and rising incomes? These response variables may seem a bit arbitrary, but they reflect a process of modeling that focused specifically on what was predictable, and where the modeling was able to effectively describe and discuss the difference between neighborhoods in Detroit, and tell a story about why such distinctions existed. The project as it is more broadly conceived looks at the success or failure of neighborhoods, and this is a set of data and insights that could be of use to city planners, people interested in revitalization, nonprofit investors in the city, and nonprofits more broadly. This project is aimed to help city leaders think about what some preconditions for different types of civic success or failure look like, localized to specific neighborhoods. The hope is to provide some insights into the distinctions between rich and poor neighborhoods in the city, and then account for those differences.

This should be seen as a starting point, which helps to distinguish one set of neighborhoods from another, and speak to these bifurcated narratives of Detroit. A somewhat more detailed follow up, which is beyond the scope of this specific project at the moment, will consider what makes a neighborhood and successful and what interventions predicted the success of a neighborhood.

Data Cleaning

For this project, I decided to focus on neighborhoods, rather than census tracts, council districts, or zipcodes, as meaningful units of measurement, or more simply, as observations in my dataset. This was primarily because they were more analytically meaningful than census tracts or zipcodes, which had no specific relationship to a local community, and because there were more of them (207) than council districts (7). Two hundred and seven is not a huge amount of observations on which to test machine learning models, but it does present a decently wide array of options. For the data cleaning section of this project, I primarily focused on taking open data from https://data.detroitmi.gov/ and the American Community Survey to specific neighborhoods. I downloaded a geojson file for Detroit's neighborhoods, and built a dictionary of neighborhood name to the polygon of the neighborhood shape. I then built a function that identified if a point was within a neighborhood or not, and if it was, I added it to a list, and at the end, I added this list back to the dataframe, providing each of these dataframes of open access data with a neighborhood column. For this part of the project, I drew extensively on the python libraries folium, shapely, and geopandas. I was also able to make use of the dissolve function in geopandas to aggregate census tracts into neighborhoods. This describes the main thrust of my data cleaning work.

As part of this broader process of data cleaning, I had to convert the various latitudes, longitudes, and addresses into workable lat, long, geometric point objects. Often, I had coordinate points, but from geojson files, and they were converted to strings, so I had to strip the coordinates of extraneous text, split them into two, convert them into floats, then zip them together, apply the shapely Point function to them, and finally, put them back in the original dataframe. In other cases, I did not have coordinates to work with, and had to geocode these addresses, using the geocoder tool from geopandas. This was a computationally costly to a prohibitive degree, and ultimately led to me having to exclude these datasets. In general, I thought geocoding would be easier than it was.

In this section, I'm going to briefly describe all the different datasets I cleaned and prepped for use in my models. In the exploratory data analysis section, I'll say a bit more about the specific categories that I imported into the large dataframe that I used to build my machine learning models.

- Libraries--the names and coordinates of all public libraries in Detroit.
- Detroit Demolitions --a dataset from January 1, 2014 to the present, which lists the names, sites, and prices of demolished buildings. One particular problem Detroit has had is with abandonments or degradation of building stock, and some agencies in Detroit (the city, the land bank, and the building

authority) all perform demolitions on abandoned or dangerous structures. Demolitions might tell us something about the nature of the neighborhood (perhaps there is lots of development or lots of abandonment).

- Business Licenses—a dataset of all business licenses issued since 2015. It does not include the type of business.
- Auction Sales—The Detroit Land Bank buys and sells vacant properties to online bidders in an auction. The dataset begins in June 2014.[2] Auctions might predict a neighborhood in decline, or perhaps show a vacant neighborhood being filled with people.
- Fire— the fire dataset includes the time, location, response time, and nature of incidents that the Detroit Fire Department responded to. This dataset spans January 2015 to July 2017. Fires are often predictive of vacant buildings or neighborhoods with dwindling populations.
- Building Permits—issued by the City of Detroit Buildings, Safety, Engineering, and Environmental Department. These range from 2010 to the present, and include new building permits as well as alterations. Building activity primarily reflects a changing neighborhood.
- Annual Inspections—Inspections of commercial buildings by the city of Detroit inspectors. These inspections range from 2015 to the present, and are scheduled for all commercial properties annually (commercial includes businesses, nonprofits, churches, government agencies, and residential buildings with five or more household units). Inspections ascertain compliance with the Property Maintenance Code, the Zoning Ordinance, the Michigan Building Code, and other regulatory codes. Certificates of Compliance are issued annually to each building, and owners must acquire these certificates, whether their building is active or vacant.[3]
- Fire Stations—a dataset of all the fire stations and their locations in Detroit and their coordinates.
- Blight—this dataset shows blight violations that have been issued property owners who have violated City of Detroit ordinances that govern how property owners maintain the exterior of their property. The dataset starts in 2004 and goes into the present.
- Childcare –a dataset of all childcare providers in Detroit.
- DDOT Bus stops—bus stops for the buses operated by the Detroit Department of Transportation. The data comes from August 2016.
- Police—locations of police stations in Detroit.
- Schools—locations of schools in Detroit.
- SMART Bus stops—bus stops in Detroit from Suburban Mobility Authority for Regional Transportation. SMART runs buses between Macomb, Monroe, Oakland, and Wayne Counties, and primarily provides the main non-private car mechanism for people to travel to downtown Detroit.
- Traffic signs—locations of all stop, signal, and yield traffic signs in the greater Detroit area. Fewer traffic signs might be a proxy for more density, or perhaps less government attention to a neighborhood.
- Improve Detroit—this dataset comes from a mobile app, which allows users to report quality of life issues like potholes, running water, and damaged street signs to City Hall, along with photos of the problem.[4] The log of issues starts in December 2014, and runs to the present.

---

[2] Due to the vagaries of rental markets and elusive owners, there are some issues with this method of disposing of property. See this article at NextCity on one case https://nextcity.org/features/view/detroit-foreclosures-tax-auction-loveland-technologies-jerry-paffendorf.

[3] For more, see here http://www.detroitmi.gov/Portals/0/docs/Permits/BSEED/PropertyM/Commercial%20Buildings%20Inspections,%20Requirements%20and%20Responsibilities.pdf

[4] http://www.detroitmi.gov/How-Do-I/Mobile-Apps/ImproveDetroit

- Parcel Point Ownership—a dataset on the ownership history of the various parcels of property in Detroit. The data goes all the way back to 1912, but is only fairly complete beginning around 1968. From this dataset, I produced a mean home price value by neighborhood dataset that starts in 1968 and goes to 2017. This dataset comes from the Assessor's Office, and includes data on the assessed value of the property, the taxable value of the property, size, height, and features of property, and its sales history.
- Census data—I drew together vital demographic, income, housing, health, ethnic and racial composition, and other data from the 2015 American Community Survey, and used dissolve from geopandas to match this data with neighborhoods. This data was represented in estimates and percentages with margins of error for each census tract, and I stripped out everything but the estimates for the purposes of the model.
- Crime data—I looked at crime stats from 2011-2014 and 2016. I drew on two datasets, which only listed major crimes, and used this as a proxy for the broader crime rate.
- Homicide data—I pulled in data on homicides from a variety of different sources, datasets from 2014, 2015, and 2016, as well as the broader crime rate dataset. I thought it was worthwhile to make homicide into its own dataset, inasmuch as homicides could be seen as a proxy for the seriousness of crime in a neighborhood, and homicides could influence the willingness of residents to stay in a particular neighborhood.

The main issue with these different datasets is that they are inescapably presentist in their presentation of Detroit; very few of these datasets extends beyond the 2010s, much less the 2000s or the 1990s, so it is difficult to tease out causal stories or extensive time series data from these different datasets. That being said, the question that drives this data analysis is the state of neighborhoods in the present time (more or less, 2015 or 2016, depending on the target variable), so the research question is what factors influence and predict particular aspects of the current state of these neighborhoods, and present factors might have strongest influence on these current situations. If I had more data, or if the city of Detroit had published earlier data, it would have been extremely helpful to have data from before the Recession, as that is often seen as a turning point in Detroit's economic history, and that would really allow me to trace some of its effects in a variety of different domains.

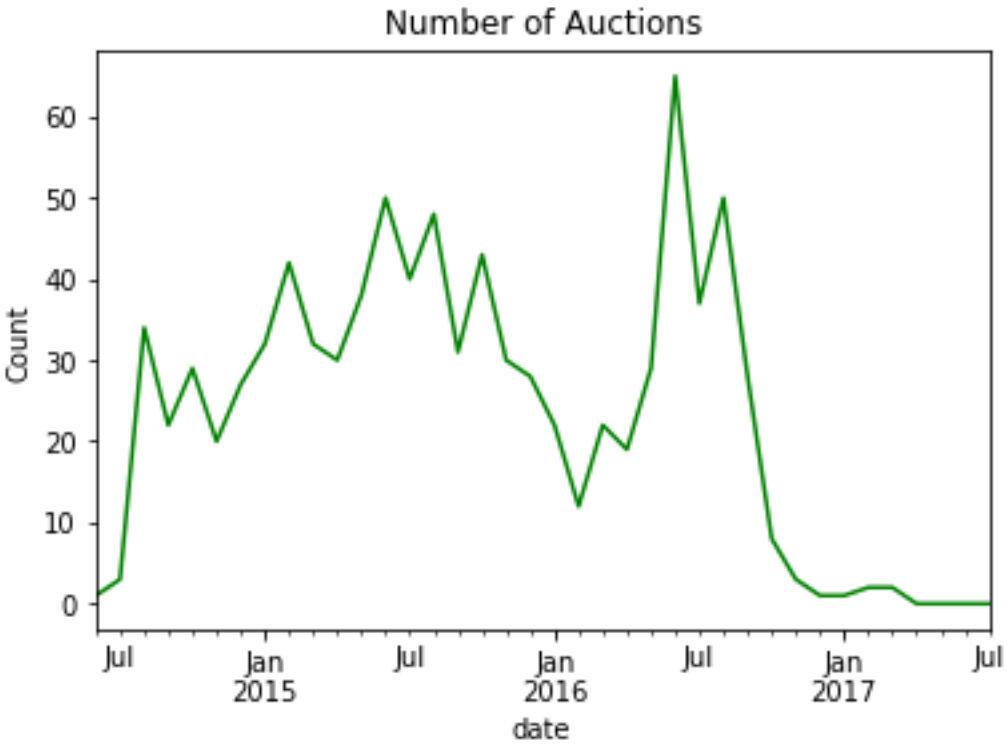
I have a lot of datasets, but the other datasets that I'd like to have would focus on two main things: work and health. For work, I'd like to know more about businesses, what their profits are, how many people they employ, who their customers and markets are, and how they relate to the general fabric of the city. For instance, it might be useful to know if a neighborhood suddenly becomes home to a series of high end coffee shops. Alternatively, do industrial areas remain industrial or do they deindustrialize? Over what time frame? How many people work in industry? From a healthcare perspective, I'd like to know more about the diseases and treatments of particular neighborhoods, particularly given that that might be useful for isolating neighborhoods that have undue environmental burdens, relative to the rest of the city.

Exploratory Data Analysis

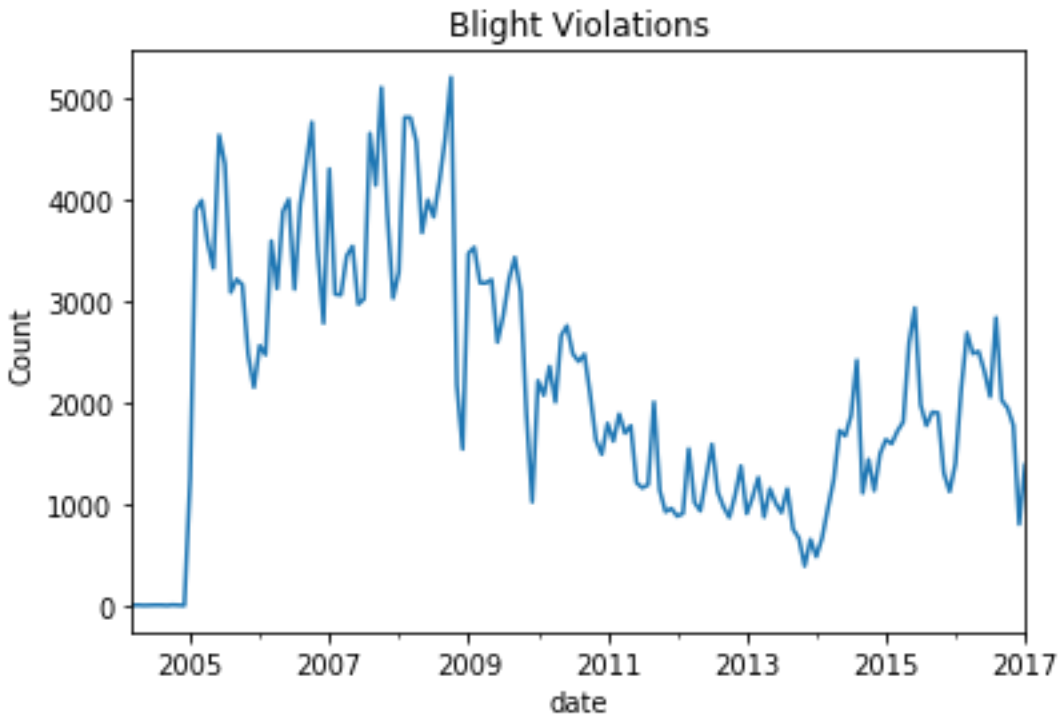
In this section, I primarily describe which features of the data I ultimately ended up incorporating into the larger dataset and why. I include some distribution charts from time to time, although the primary goal here was to find statistics that were meaningful for each neighborhood from these datasets. I also did some time series graphs where relevant. For each dataset, I picked out some elements that generalized well to all the neighborhoods and were meaningful.

- Auctions—from the auctions dataset, I took variables for purchaser type (Investor/Homeowner), total number of auctions, and average sale amount. I also made a time series graph of total auctions

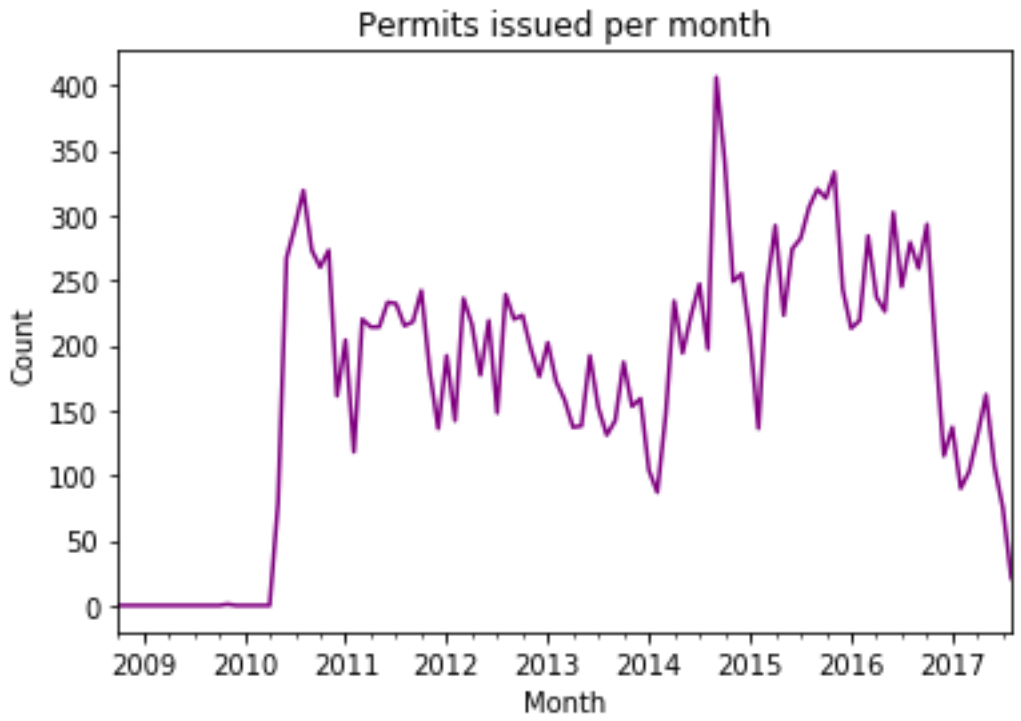
over the course of the dataset, which suggested that the level of auctions had remained fairly constant, with a brief spike in 2016.



- Blight violations—from this dataset, I took variables for the average fine amount levied for a blight violation and the number of blight violations total per neighborhood. I also noticed a dramatic difference in the number of blight violations issued before and after the recession of 2008-2011, as shown in the graph below. This seemed to me to probably represent a somewhat different enforcement regime for blight violations, pre and post recession. On the basis of that distinction, I decided to group blight violations by if they had occurred before or after the recession, using the column names blightviolations2000s and blightviolations2010s to represent these differences in the sheer number of blight violations that were issued.



- Building permits—this dataset includes all building permits issued between 2008 and 2017. I pulled a variety of columns from here for my neighborhood dataset, including length, the average length of construction, the number of new buildings in a neighborhood, the average size of parcels, the average estimated cost, and the average floor area. I also turned the type of building constructed, which included identifications of buildings like multifamily home, single family home, ATM, and restaurant into a dataframe, and joined it with the larger dataset. Then, for each year represented in the dataset, I made a category with the year and the count of buildings permitted. This general count of building per year includes all building permits issued in a neighborhood, and I felt that that was appropriate, because changing residential patterns in a neighborhood are not wholly the result of new construction, however, new construction seems important enough that it deserves its own category. One interesting feature I noticed in this dataset is that there is a huge spike of building permits issued in 2014, which is when downtown Detroit really started being built up.



- Annual Inspections dataset—I made a dataset of all the action descriptions, which reflect further actions suggested by inspectors, broke them down by neighborhood, and joined them to the main dataframe. I also made a column for the number of inspections per neighborhood per year (2015, 2016, 2017).
- For the datasets Bus Stops, Childcare Providers, Fire Stations, Libraries, Police Stations, SMART Bus Stops, and Traffic Signs (although this was really three columns, signal sign count, stop sign count, and yield sign count), I merely made a count column for each dataset, and turned it into a column.
- Improve Detroit dataset—I made a dataset of each type of issue (Graffiti, Trash, Dead Animal Removal) and grouped them by neighborhood, so that each issue is a column of its own in the dataset. I also made an avg response time length by neighborhood and added that to the dataset, as well as a count of Improve Detroit Tickets and the average ratings of the service. I also noticed that Improve Detroit Tickets tend to be cyclical and fall off in the winter time.
- Property Values dataset—for this dataset, I filtered it to focus on buildings that were taxable. I then made columns for the median/mean assessment value (which is the price that the City of Detroit thinks a property is worth, and it grows at a set rate, even if the neighborhood or the market might value a house more. This can change if a property is sold, as the assessment value and the actual value become realigned), cost per square foot, average size of property (in acres), the average year that houses in the neighborhood were built, and a larger dataset on the salesprice.
- For the two major crime datasets, I made a total crime per year for each dataset, and then broke it down into categories per year (and there were six total). Homicides had their own dataset.
- Business licenses—on exploring the data, I included four columns, total business license per neighborhood and a breakdown of this data, which counted businesses permitted in 2015, 2016, and 2017.
- Demolitions—for this dataset, I did much the same as above, breaking it down into total demolitions, and then demolitions in 2014, 2015, 2016, and 2017.

Modeling and Preprocessing

In this section, I'm briefly going to describe the various steps I undertook to model these datasets. First, having sorted all of this data, located it by neighborhoods, and provided adequate summary statistics from each dataset, I concatenated it altogether into one large dataset. In total, I had 712 columns on 207 different observations (the vast majority of this was census data). From this dataset, I extracted a variety of dependent variables, ranging from income to businesses permitted to new buildings constructed, and put them in a separate list. I dropped all strings and indexes from the dataframe, replaced NaNs (of which there were many) with zeros. This was not ideal, since the data was at times incomplete, but usually a NaN meant that something hadn't happened to the same degree in a particular neighborhood, so I think that was an appropriate means of handling the issue.

Having appropriately cleaned my dataset, I created a set of functions to quickly implement all of the different models that I wanted to try on the data. Since there was no set of labels to classify this dataset on (I suppose I could have generated them, but this seemed to be about relative progress and prosperity as opposed to comparative progress), I only used regression models, including Lasso, Ridge Regression, Random Forest Regression, Decision Tree Regression, Gradient Boosted Regression, and Support Vector Regression. In cases where it seemed beneficial, I also implemented feature selection, using a function to pull the non-zero coefficients from Lasso, and then feeding that back into a linear model to see if feature selection improve performance notably. For each function, I also printed out the RMSE on the training and test data, and then used this comparative performance to evaluate the model. RMSE was my main scoring mechanism, and I used it to compare model performance, and tune the models effectively.

In these different functions, I included tuning code (primarily GridSearchCV), which I modified when I called each function. The parameters I tuned for each model were as follows:

- Lasso and Ridge—alpha values (needed to be higher for my dataset) and a tol value of 0.1, as the models sometimes had difficulty optimizing
- Decision Tree Regression—max_depth, max_leaf_nodes, and min_samples_leaf
- Support Vector Regression—C, epsilon, and gamma (and I used an rbf kernel)
- Random Forest Regression—max_depth, max_leaf_nodes, and min_samples_leaf
- Gradient Boosted Regression—max_depth with a learning rate of 0.1

Some brief observations on model performance, which will shape the results. The best models I built were from Lasso and Gradient Boosted Regression. Random Forest worked from time to time, but SVR and Decision Tree Regression were fairly superfluous. Ridge Regression performed fairly well, but I think the data was not sufficiently clean for it to work, as it generated a series of scipy linear algebra errors, including a warning that the data might not be accurate. For that reason, I ultimately ended up not using any of the Ridge Regression models here. Lasso and GBR were the most effective models, I think somewhat intuitively, because there were so many columns in this dataset. So Lasso set many at coefficients to zero, and was able to produce a more useful model through feature selection, reducing the dimensionality of the dataset, while GBR kept running new models on the residuals, also performing an internal form of feature selection. Decision Tree and Random Forest Regression also seem to have been less potent here, as there was little categorical data in this dataset, most of it was numeric, and served as summary statistics from other Detroit datasets.

Models

I built a lot of models, and modeled a lot of features. I tried to stay focused on aspects of this what were useful and interesting. I may have failed a little bit.

Total Buildings Permitted: One issue that was of interest to me is how fast are neighborhoods growing, and what factors influence that process? I used Total Buildings Permitted as a dependent variable, which is necessarily a bit flawed, as it includes a variety of different types of work. I tried to do something similar with new buildings alone, but the numbers were so low for most places that the models were highly flawed (RMSE of around 40, mean new buildings were 8). However, it seemed to me that large scale renovations that require a permit would suggest a sufficiently growing neighborhood in aggregate. For this, I used a lasso model with an alpha of 1, which had the following performance on the training and test data:

Total Buildings Permitted Mean: 236.9

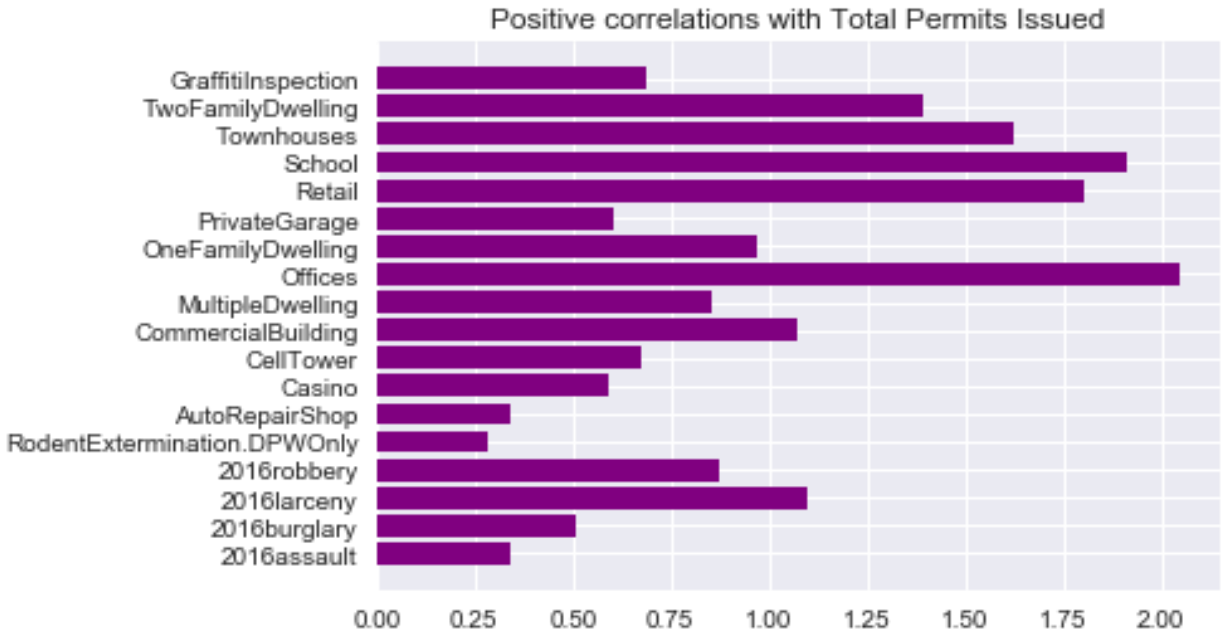
Best alpha: 1.0

Number of features used: 313

Root Mean Squared Error on train data: 1.9195287277293813

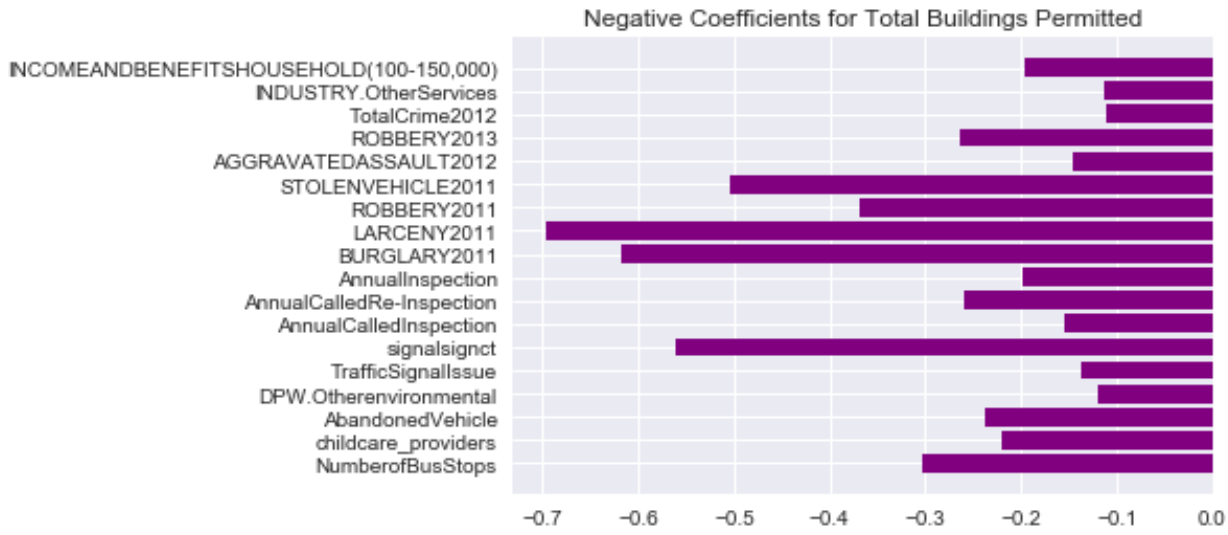
Root Mean Squared Error on test data: 43.04753879564554

This model is reasonably robust, having an average error of about 43, which is a significant error, but not too significant (and indeed, some of the models I tried had errors that were 3 or 4 times the mean). The point is that this has reasonable predictive power concerning the total buildings permitted in a neighborhood.



This chart looks at positive coefficients and their impact on Total Permits Issued. Large impacts come primarily from different types of buildings constructed, such as Cell Towers, Casinos, Multiple Dwellings, TwoFamily Dwelling, Schools, Retail and Commercial Buildings. These building types suggest that total buildings permitted serves as a useful proxy for expanding construction. At the same time, neighborhoods in which such new buildings are put up are those that tend to fare better than other neighborhoods. Intriguingly here too is a series of factors related to crime in 2016, perhaps suggesting that

crime rates may have risen in neighborhoods with more buildings permitted in 2016 as population patterns change.



Negative coefficients primarily have to do with crime, particularly in 2011, 2012, and 2013, which are crime stats that could have influenced siting and permitting decisions around 2016 and 2017, which is where most of the permitting data comes from. Prominent here as well is the census data point households with a value of 100,000 to 150,000 dollars, i.e., richer people. Richer people often tend to oppose development for a variety of reasons, and that could be an important predictor here as well. I am not sure what to make of the role of inspections here; inspections are annual events that each business and large residential building in Detroit has to have done, and perhaps what is suggested here is that more buildings are not necessarily being built in neighborhoods with lots of preexisting building and density, but are instead being constructed in neighborhoods where new buildings are going up, and there aren't a lot of businesses to be inspected. The number of bus stops may speak to a similar phenomenon, and childcare providers as well.

Mean Household Income: The next model looks at what factors correlate with a higher mean household income in a neighborhood. For this model, I used feature selection, from the lasso model, and put its nonzero coefficients through a selection process to build a feature selected dataframe, and then trained a Gradient Boosted Regression model on this dataframe. The lasso model had an alpha of 1, and selected 632 features (with an error on the test data of 20,000). The GBR model used 4 as its max_depth_parameter, and produced the following results:

Mean Household Income: 42321.783

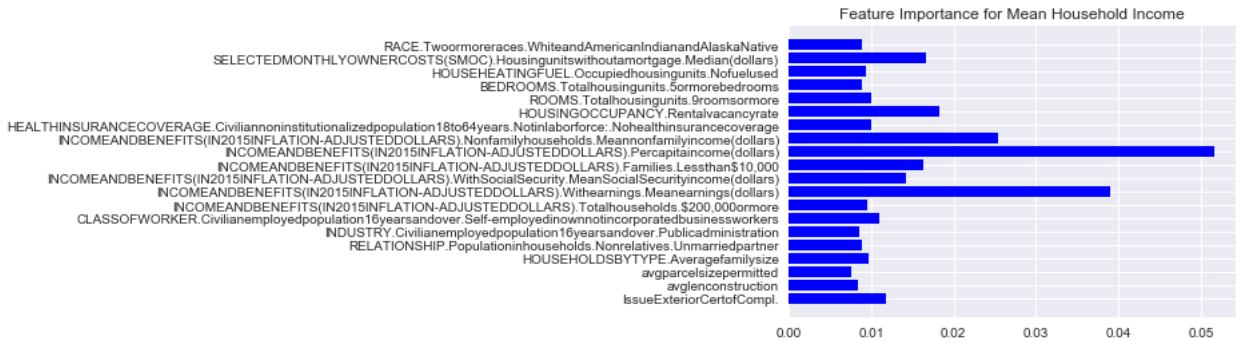
Median Household Income: 39926.0

Root Mean Squared Error on train data: 0.29

Root Mean Squared Error on test data: 1193.188

While 1193 is relatively high, especially at the lower ends of the spectrum, this model performs dramatically better than pretty much any other option. Further, given the actual numbers under discussion, 1193 is relatively small, something 2.8% of the total income in question here. GBR does not have coefficients, but rather, feature importance, which is a measure of the degree to which a feature is able to

influence a model. I filtered these features at an importance of 0.075, which might be a bit low, but provides some useful and interesting context, and is expressed in the following chart:

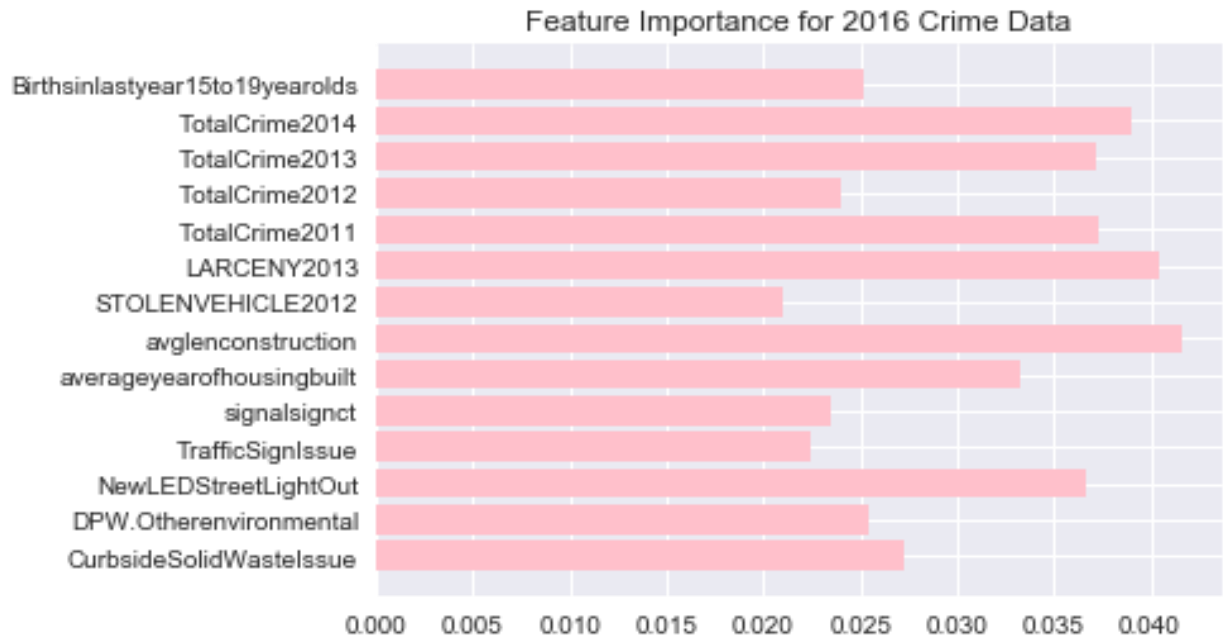


This is primarily drawn from census data, and the largest coefficients are essentially proxies for incomes, things like Per Capita Income, Mean Earnings, Families making more than 200,000 dollars and less than 10,000 dollars, which would quite obviously have a strong impact on the mean income of a neighborhood. Other less obvious factors also serve as interesting proxies for income size, such as the vacancy rate of a neighborhood—it would seem that higher vacancy rate means lower mean household income. Rooms are also a good proxy for the size of a house, 5 or more bedroom and 9 rooms or more, also probably signify the house of the well to do. Average parcel size, average family size, and the average length of construction also suggest differing neighborhoods, and employment data highlights the importance of public administration and the self-employed, perhaps having a negative correlation with mean household income. In some ways, much of this is obvious, but the particular contours are a bit surprising and what played a specifically important role.

2016 Crime Data: This model concerns crime data in 2016, and looks at what factors correlate with it. This crime data is separate from homicides, as homicides are much rarer than crime in general. For this model, I used lasso with an alpha of 5, and 131 features for feature selection, then took this dataset and put it into a Gradient Boosted Regression Model with a max_depth_parameter of 2. The vital statistics are as follows:

Crime Mean: 150.1352657004831
Root Mean Squared Error on train data: 2.8323436921143834
Root Mean Squared Error on test data: 35.45128560571354

I filtered the feature importance at 0.02, which produced the following chart below:



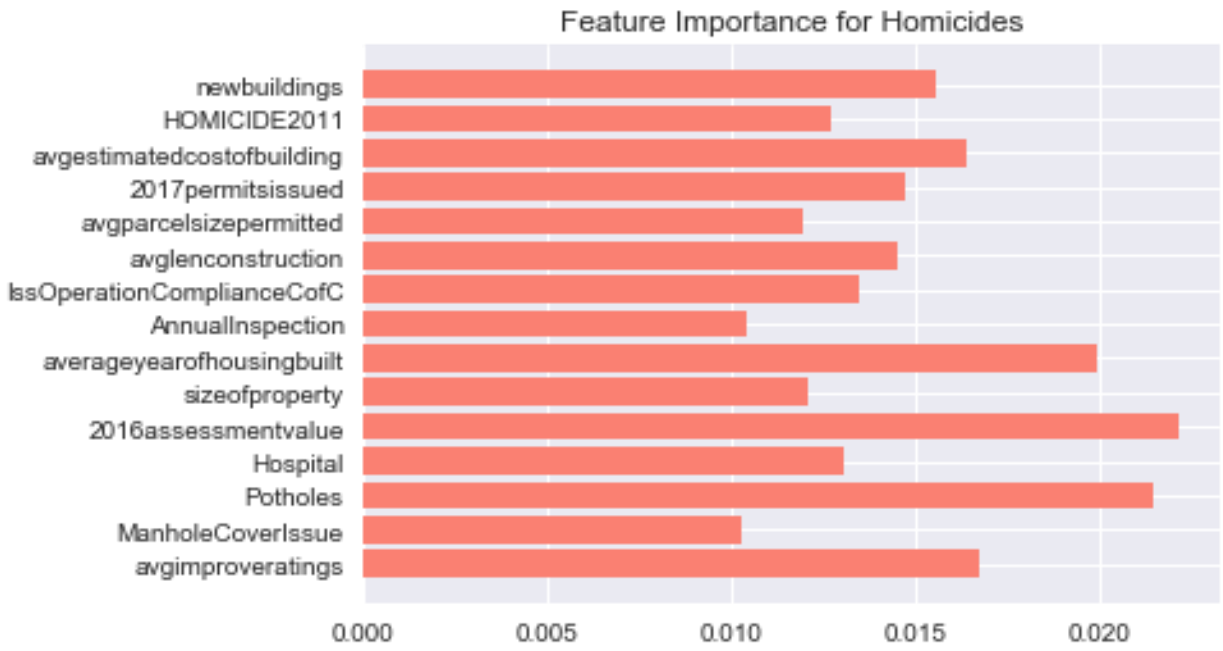
It should be briefly noted that the Open Data Detroit website lacks any data for 2015, so hence the reason it does not appear in this chart. What pops out clearly here is that crime in 2016 is highly influenced by where crimes occurred in the past, and that Larceny and Stolen Vehicles seem to have some influence on this dataset. This is helpful from a public policy perspective, as it suggests where to deploy resources. Significant here also is teenage pregnancy, although it is unclear if this is a symptom of areas with high crime or if it is causative. Average year of housing built probably has an inverse relationship with crime data (i.e, the newer it is, the less crime there is), although it is not obvious from the model. I would assume that average length of construction functions similarly, as shorter construction times lead to less crime, although again, this is not clear from the model. Finally, there are a series of issues related to Improve Detroit or other citations and complaints that predict crime, such as Curbside Solid Waste Issue, Traffic Signs, or New LED Street Light Out, all of which could play a role in predicting potential crime levels in a neighborhood, and make these particularly important issues to watch out for, as an increase in them might bode an increase in crime. It is possible however that these are accompanying symptoms as high crime neighborhoods have them as well. But regardless of the actual relationship, from a macro policy perspective, these features seem useful to at least review and track.

I also built a model to predict the total number of homicides in a neighborhood in Detroit. This is somewhat related to crime above, but actually involves a surprisingly different set of feature importance. The model I used was gradient boosted regression, and I trained it to predict total homicides in a particular neighborhood, rather than just a specific year, as I figures this had higher odds of success.

Average Homicides per Neighborhood: 2.855

Root Mean Squared Error on train data: 0.005769891832668183

Root Mean Squared Error on test data: 5.783756023762857



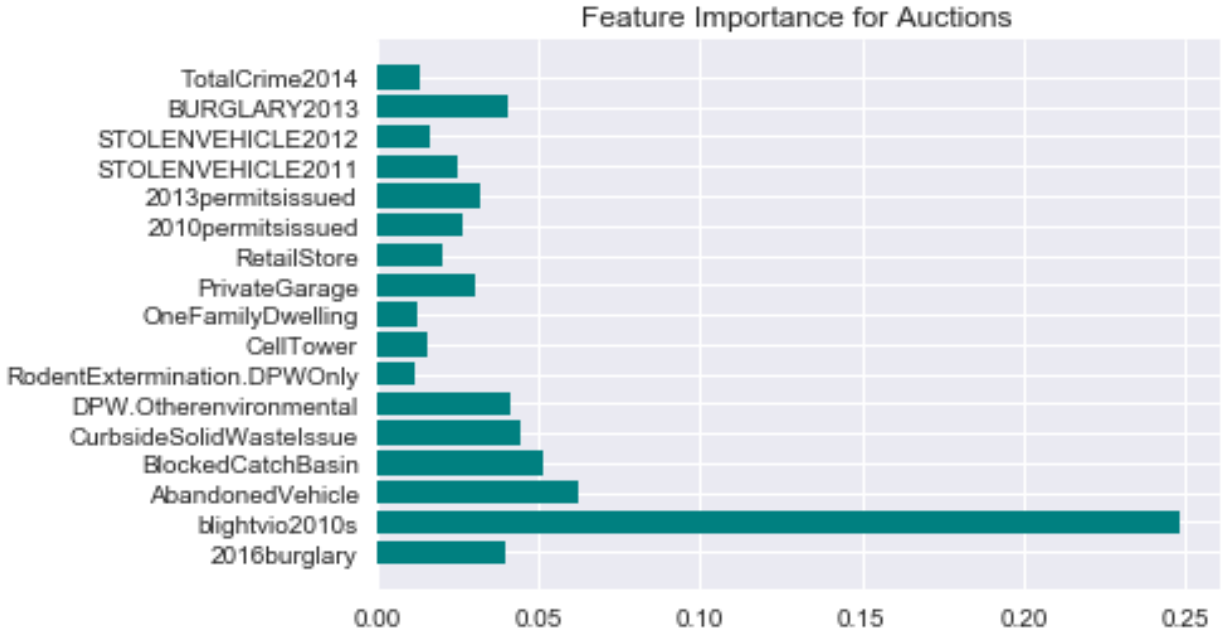
I actually find this model rather perplexing, as you might expect homicides to have some connection to crime. Instead, it tracks 2011 homicides, and things like assessment value, average year of housing built, average length of construction and average parcel size permitted. It also seems connected to some issues like Potholes and Manhole Covers, and Hospitals (perhaps proximity to hospitals influences reporting of homicide crimes?). It might be that a lot of these features are proxies for rich/poor neighborhoods, and that is the core division between these areas; homicides don't happen in wealthier neighborhoods to the same degree. Annual Inspections and new buildings could function similarly, as they reflect areas with more businesses/more development.

Total Auctions: A good measure of communal stability and lack of turnover is the number of auctions in a particular neighborhood. Auctions are put on by the Land Bank of Detroit, and are thus, often a good measure of displacement. Auctions are in their own manner a measure of displacement, as they are often properties that have blight violations or have fallen into disrepair. There are also reports of unethical uses of blight violations to drive particular tenants out of neighborhoods. Auctions are a means of disposing of these properties for the land bank. Increases in auctions mean more houses sold off by the Land Bank, and less stable neighborhoods, i.e. neighborhoods that for whatever reason are not thriving. For this model, I used a random forest regression, with a max depth of 4, max leaf nodes of 8, and min samples per leaf of 2. The accuracy was as follows:

Mean Auctions per Neighborhood: 4.4975845

Root Mean Squared Error on train data: 4.355330534999539

Root Mean Squared Error on test data: 8.54984426243721



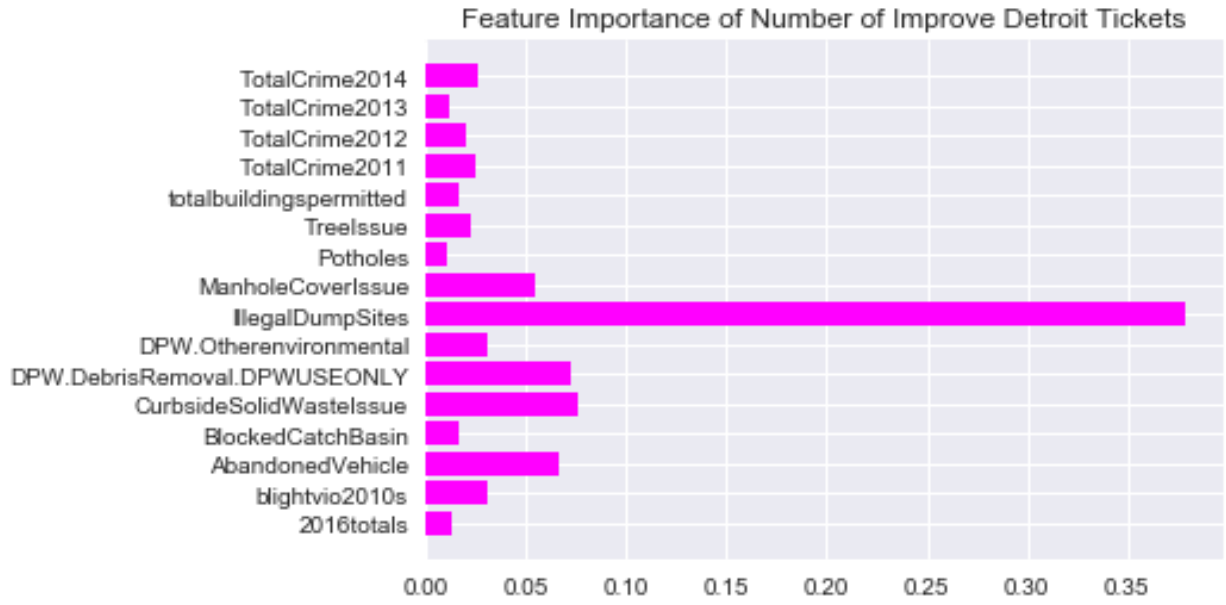
This model performs decently well on the dataset, and has an error that is not too far outside the mean of the observations here. It can accurately predict for the most part the number of auctions that a neighborhood will have. One feature of this model that seems highly important is the strong correlation between the number of auctions a neighborhood has and the number of blight violations that a neighborhood has. Blight violations often lead to auctions, as enough violations lead to the forfeiture of property rights, and then the house is auctioned off. This tight connection should also be worrisome to policy makers, as blight violations might be causing displacement from neighborhoods. Other things here like Blocked Catch Basin, Abandoned Vehicle, and Rodent Extermination all seem like lower level correlates to blight violations. Crime also appears to be an issue here, and perhaps particular types of crime in neighborhoods correlate to lower property maintenance or abandoned housing. Permits issued and the types of stores may correlate in the opposite direction, perhaps being things that lead to more neighborhood stability, although it is difficult to know with a Random Forest Regression model.

Improve Detroit Tickets: Another worthwhile model predicts the number of Improve Detroit Tickets that are submitted in a neighborhood. Improve Detroit is a service/app that allows users to report issues in their neighborhood, which the user selects from a menu of quality of life issues and they are filed in a database alongside a picture. This tool helped city staff quickly identify and fix a problem. My intuition for these tickets was that they would reflect areas with more smartphones or with wealthier residents, rather than reflecting the more general experience of the city as a whole. Thus, I thought that the number of tickets placed or requested might have some direct relationship to the more successful neighborhoods in the city, rather than the specific problems that they were reporting. The model I built for these tickets was a Random Forest Regression model with a max depth of 7, max leaf nodes of 13 and minimum samples per leaf of 2. For this model, I first used lasso for feature selection, then put the new feature selected dataframe into the random forest regression algorithm.

Average Number of Improve Detroit Tickets: 884.6763285024155

Root Mean Squared Error on train data: 474.6401565768995

Root Mean Squared Error on test data: 145.6483962647053



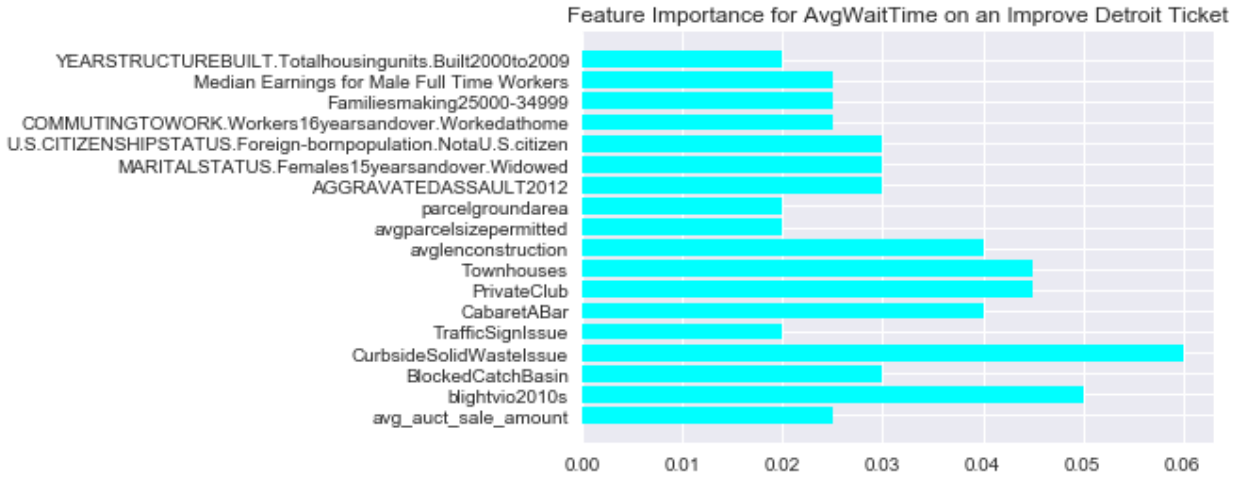
So this intuition that Improve Detroit tickets were only a symptom of richer neighborhoods turned out to be wrong, as they show no strong correlation with well to do neighborhoods. Instead, they tend to reflect actual issues (also from the Improve Detroit database). These issues and the use of the Improve Detroit process appear to be distributed evenly, as there is no real separation between neighborhoods on other grounds here. The strongest connection of a ticket to a neighborhood has to do with Illegal Dump Sites, which have the strongest feature importance here. I am not sure if this is just the result of a strong signal in some neighborhoods, but it seems from this dataset that illegal dumping causes a very strong reaction. Other issues that elicit similar reactions are Blocked Catch Basin, Manhole Covers, Trees, Abandoned Vehicles and Curbside solid waste. Another predictor of the usage of the Improve Detroit database is the level of crime in a neighborhood (including 2016totals). It seems likely (although again it is random forest) that higher crime neighborhoods might have more necessity for Improve Detroit reports and that there is some correlation between some of these issues and Improve Detroit tickets. Blight violations are also a useful feature in this model, as they too seem to correlate with more Improve Detroit tickets. Total buildings permitted might suggest that in areas with more people, there are more tickets. Alternatively, there might be construction issues in those areas, which drive things like reporting illegal dump sites or other potential violations.

I also built a model for the average wait time of an Improve Detroit ticket, suspecting that this could potentially tell us something about the different fate of neighborhoods. For this model, I used gradient boosted regression, and formed this model by subtracting the time of completion from the time that the ticket was posted, grouping this data together by neighborhood, and then averaging it (in days).

Average Wait Time on Improve Detroit Ticket: 18 days

Root Mean Squared Error on train data: 2.864260201625337

Root Mean Squared Error on test data: 8.50389998120845



It is here that we begin to see some economic differences between the response time. It should be cautioned first that the feature importance of the coefficients in this model is quite low, and indeed, only the Curbside Solid Waste Issue has a feature importance of more than 0.05, but we can perhaps speculate on some of the ways that neighborhoods are separable, based on this dataset. First, a set of strong predictors of wait time have to do with the actual issues themselves, such as Curbside Solid Waste or Traffic Sign, suggesting that it might be simpler or more difficult to respond to a particular issue. Blight violations might also be correlated here as well. Median earnings, the income of specific families, widows, and non US citizen populations probably correlate with less well to do neighborhoods. Meanwhile, avgparcelsize, parcelgroundarea, and avgauctsaleamount tend to be measures of the wealth of a neighborhood. Townhouses, Private Clubs and Cabarets/Bars are probably symptomatic of wealthier neighborhoods. My operating assumption would be that poorer neighborhoods take longer to be served, although we cannot know this directly from the specific model that I am using. So while the usage of Improve Detroit tickets is not segmented by income, the wait time for the resolution of an issue is.

Conclusions

I have focused on six target variables, crime/homicide, Mean Household Income, Total Auctions, Total Buildings Permitted, Total Improve Detroit Tickets and average wait time for Improve Detroit Tickets. What I have endeavored to show is the significant difference in outcomes between different neighborhoods in the city of Detroit, and to track some of the ways that the city is succeeding or failing in servicing its neighborhoods. What I have found overall is that there are strong divides between succeeding and failing neighborhoods, and these primarily tend to be proxies for wealth of residents; areas with wealthier residents tend to do better overall in these metrics. The success of a Detroit neighborhood on these metrics depends heavily on its level of income, but this process is all intertwined.

For instance, higher crime rates lead to less total buildings permitted. Past crime is a high predictor of present crime in a neighborhood. The model for 2016 crime data also seems to suggest that income plays a key role in minimizing crime, there are several proxies for higher income neighborhoods in the dataset. The average wait time for Improve Detroit tickets also seems to have several proxies for income, yet the actual Improve Detroit tickets themselves tend to be uniformly adopted, and the total generated by a neighborhood seems to have little to do with income, and more with the specific issues confronting a neighborhood. Mean income prediction of a neighborhood primarily seems to be a function of its income or various proxies for it such as house size or more than five rooms.

These findings would all seem to support the narrative that the recovery of Detroit has been staggered, concentrated in some particular neighborhoods, rather than in other neighborhoods. The key takeaway here is that the mean income of an area goes a long way towards predicting the development patterns, crime patterns, and how long neighborhoods might wait for a particular service (not all services—this is just one example). This would seem to suggest that the benefits of the Detroit renaissance accrued to a few areas of the city, and that some wider and more redistributive framework should be adopted to help other parts of the city along. Some thoughts on this might include:

- Incentives to move for residents of some neighborhoods
- Targeted city investments in specific neighborhoods that are not doing as well (policing, building projects, public buildings and services)
- Increased transit options to less desirable neighborhoods to incentivize movement into them

Detroit is faced with a rather difficult problem here, as it is a vast city, and the population still does not even come close to matching the current size of the city. Given that, it has far more space to cover and offer services, and that seems to be a major reason for the disparity in outcomes between these different neighborhoods.

All of this is ultimately a bit unsatisfactory, as it is fairly clear, it's a set of conclusions that we might expect. Rich neighborhoods do better on these metrics than poor neighborhoods, and that's fairly clear for a variety of reasons. There is not a lot to do there, besides strong redistributive policies. What is more interesting is attempting to think about the specific policy outcomes and possibilities that could play a role in making the city of Detroit potentially more equitable, or more simply, what are the differences between rich and poor neighborhoods, and what specific policies can be derived from these models?

- **Blight violations need reform, they have far too strong an influence on auctions.** Blight violations can potentially be a useful revenue source for the city or a means of controlling what happens on properties, but the degree to which they contribute to auctions and displacements is highly suspicious and needs to be looked at.
- **Auctions are also connected to crime data and other issues such as Blocked Catch Basin, Abandoned Vehicle, and Rodent Extermination**. These kinds of issues might make it possible to classify and monitor the risks of auctions in neighborhoods, and eventually intervene in the process.
- **Some types of buildings seem to correlate strongly with more building permits issued.** Buildings like schools, retail, townhouses, and casinos all seem to create a virtuous cycle of building; building them leads to the creation of more building permits (perhaps for the buildings themselves?), but also potentially for the neighborhood. Thoughtful siting of these types of buildings can help kickstart that cycle in other parts of the city.
- **Specific types of crime and issues seems to have outsized predictive power.** For instance, larceny and stolen vehicles, as well as curbside solid waste, appear to strongly correlate with crime data. It might be worth looking at this data and these correlations in some more detail.
- **Homicides seem to be predicted by different features than crime more generally.** While crime data has a host of predictors, most of the homicide data has to do with the distinctions between a wealthy neighborhood and a poorer neighborhood (of course, this is up for debate, because homicide data is ultimately so much smaller and noisier than crime data).
- **Illegal Dump Sites have the highest feature importance for Improve Detroit Tickets.** A spike in Improve Detroit tickets in a neighborhood is likely to be caused by an illegal dump site. This may be because it takes so long to clean up? Or because they are so ubiquitous? But tackling this issue swiftly will really help the city eliminate redundant Improve Detroit Tickets.

- **Improve Detroit tickets appear to be distributed fairly equitably, and not concentrate in neighborhoods based on income.** The city is doing a good job with this service.
- **Curbside Solid Waste and Blight Violations are the strongest predictors of average wait time.**
- **Average wait time generally seems to correlate with income.**