Vacancy and Crime for City of Rochester

Yaxin Liu, Peiran Chen, Xiangyi Liu, Maralmaa Erdenebat
ROC Miners, DCS 383W
Goergen Institute for Data Science University of Rochester
{yliu139, pchen22, xliu84, merdeneb}@u.rochester.edu

1. INTRODUCTION

Crimes committed in the United States in both violent and property categories have fallen by 3.1 and 5.6 percent respectively in the first half of 2019 [1]. This trend is also true for the city of Rochester and since 2011, there has been a decline in the number of overall crimes committed. Despite this favorable trend, Rochester's crime index is 7 out of 100, with closer to 100 being safer, which gives Rochester a low safety grade. The city also experiences more crime per 1,000 residents compared to the national median [2]. Due to these statistics, the city of Rochester is furthering its efforts to mitigate the crime rate by identifying possible factors that could be significant.

Vacant properties are those that have been abandoned by owners and are identified by the city when tickets, violation bills and regular property payments are not being paid. Our sponsor, the city of Rochester, believes these properties can have a negative effect on the neighborhood, as unsecure and poorly maintained structures and yards could be sites of crime. Therefore, it operates an aggressive policy of owning vacant properties that can be considered for demolition or resell as well as those privately owned [3]. Apart from studying the relationship between vacant structures and crime in the city, we wanted to explore other factors that could contribute to crime.

An abundance of literature has been written on factors affecting crime over the years. One is a recent paper written by John M.A. Bothos and Stelios Thomopoulos [4], that explores past research as well as their own findings on dynamics that factor into crime. Bothos et al. consider school dropout rate, below poverty percentage, law enforcement personnel to arrests percentage as well as short-term unemployment, to be factors that hold significance in crime in the community. As such, we consider similar variables in addition to the number of vacant structures for the city of Rochester.

By defining these factors, we aim to create a model that predicts the reduction in crime rate from already proposed demolitions and ultimately recommend a data-driven process that prioritizes future demolition considerations.

2. DATA DESCRIPTION AND PREPARATION

2.1 Dataset Description

The whole project progress can mainly be divided into two parts. During phase 1, the datasets - Vacant Structures, Demolitions, and Crime were provided by the City of Rochester and, during phase 2, the datasets - Population, Education Attainment, Poverty Status and Unemployment Status and Total Population were provided by US Census Bureau Data, American Community Survey. For phase 2, with the difference in time window for the selected Census Data and the remaining datasets, we used the datasets from 2010 to 2018 to predict the data for 2019 and 2020, the detailed procedures will be further discussed in section 4.1.1 Data Forecasting.

Phase 1

Vacant Structures Dataset [5]

The dataset is provided by the city of Rochester, Department of Neighborhood and Business Development. It has 1826 records of vacant structures in the Rochester area. For each record, it included 25 features of the structure, for example its address, owner, vacant date, etc.

Demolitions Dataset [6]

The dataset is provided by the city of Rochester, Department of Neighborhood and Business Development. It has 336 records of the demolished structures in the Rochester area. For each record, it included 28 features of the structure, for example its address, whether its owned by the city, its demolishing date, etc.

Crime Dataset [7]

The dataset is provided by the city of Rochester, Rochester Police Department Open Data Portal. It has 95945 records of crime cases occurring in the Rochester area. For each record, it included 40 attributes of the crime case, for example the latitude and longitude of the location of the crime, its occurred time, crime type, etc.

Phase 2

Education Attainment Dataset [8]

The dataset was extracted from the Educational Attainment Dataset, ACS 5-Year Estimates Subject from American Community Survey.

Educational attainment refers to the highest level of education that an individual has completed. According to the Census Bureau, the data was collected by asking the question: “What is the highest degree or level of school this person has completed?” The response categories include:

- No schooling completed

- Nursery school
- Grades 1 through 11
- 12th grade—no diploma
- Regular high school diploma
- GED or alternative credential
- Some college credit, but less than 1 year of college
- 1 or more years of college credit, no degree
- Associates degree (for example: AA, AS)
- Bachelor's degree (for example: BA. BS)
- Master's degree (for example: MA, MS, MEng, MEd, MSW, MBA)
- Professional degree beyond bachelor's degree (for example: MD, DDS, DVM, LLB, JD)
- Doctorate degree (for example, PhD, EdD)

We did some calculations with the data to build the model, and it will be explained in section

Poverty Status Dataset [9]

The data was extracted from Poverty Status In The Past 12 Months Of Families, ACS 5- Year Estimates Subject from American Community Survey.

According to the Census Bureau, the data was collected by asking participants several questions about the funds a person receives from various sources to create statistics about income, earnings, and poverty. Sample questions are:

(1) Report the amount of wages, salary, commissions, bonuses, or tips from all jobs before deductions for taxes, bonds, dues, or other items for the past 12 months.

(2) Report net income (self- employment income from non- farm businesses or farm businesses, including proprietorships and partnerships) after business expenses for the past 12 months.

The rest of the questions asked can be found through Census Reporter. [10]

The data column we used for building the model is: Estimate!!Percent below poverty level!!Population for whom poverty status is determined.

Unemployment Status Dataset [11]

The data was extracted from Selected Economic Characteristics, ACS 5-Year Estimates Subject from American Community Survey. According to the Census Bureau, the data was collected by asking participants several questions about the individuals' employment status, whether the individuals are employed, or not employed or retired, and about their occupations and industries. Sample questions are:

(1) For whom did this person work? Name of the company, business, or other employer. If now on active duty in the Armed Forces, mark X of the box. And print the branch of the Armed Forces.

(2) What kind of business or industry was this? Describe the activity at the location where employed. (For example: hospital, newspaper publishing, mail order house, auto engine manufacturing, bank)

The rest of the questions asked can be found through Census Reporter. [12]

The data column we used for building the model is: Estimate!!Unemployment rate!!Population 16 years and over.

Total Population Dataset [13]

The data was extracted from Demographic and Housing Estimates, 5-Year Estimates Subject from American Community Survey. The dataset gives detailed statistical information about aging and housing on the tract level.

The data column we used for building the model is: Total!!Estimate!!Total population.

Integrated dataset

With the dataset provided by the city of Rochester, integrating with the census data, we build our finalized dataset including the following features:

- Address of the vacant structure
- Zip code

- Latitude
- Longitude
- Block number
- Tract number
- Number of vacant structures in the tract
- Population of the tract
- Population of the tract with lower than high school education attainment
- Unemployment rate of the tract
- Poverty rate of the tract
- Count of crime cases before and after demolition

2.2 Data Preprocessing

2.2.1 Data Cleaning

(1) 18 records from the Demolitions dataset were deleted because for these records only part of the property (i.e. garage only, front only) got demolished.
(2) 8 records from the integrated dataset were deleted due to the missing values of count of crime cases before and after demolition.

2.2.2 Feature Engineering

We created new features for the integrated dataset from the datasets used in phase 1 and phase 2.

(1) For the feature "Population of the tract with lower than high school education attainment", we sum up the Population 18-24: less than high school graduate and Population above 25: grades less than 9th grade and 9-12th grade with no diploma from the original Educational Attainment dataset.
(2) For the feature "Crime before and after demolition", we calculated the number of crime cases from the Crime dataset, with a time window of 1 or 3 months before and after the demolition date within 1 mile radius from the demolition address.
(3) For the feature "Number of vacant structures in the tract", we calculated the number of vacant structures in each tract from the Vacant Structure dataset.

3. EXPLORATORY DATA ANALYSIS

Before building the model to predict the effect on crime rate after demolition, and further recommending a data-driven process for prioritizing demolitions, we first need to have a deeper understanding of the data. In this section, we will focus on the overall pictures of the datasets we worked on and the correlation among features.

3.1 Overview of Datasets

Before further investigation in the correlation between features, we mentioned in the last section that we would like to obtain an overall picture of the datasets we would be working on.

Fig.1 shows that most of currently existing vacant properties had a vacancy history of 2-3 years. However, 100 (5.5%) of vacant properties have been vacant for over 10 years, and 5 (0.2%) of vacant properties have been vacant for over 20 years. Fig.2 indicates that there is a decreasing trend for crime from 2011 to 2019, and we aim to explore what was the cause of the trend and how we could keep it to help improve the safety of our neighborhood.

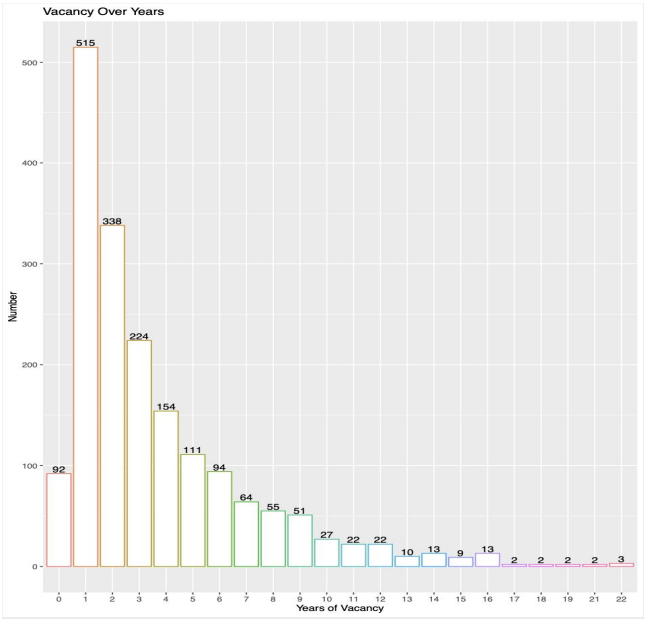


Fig. 1 Overview on Existing Vacancy Dataset

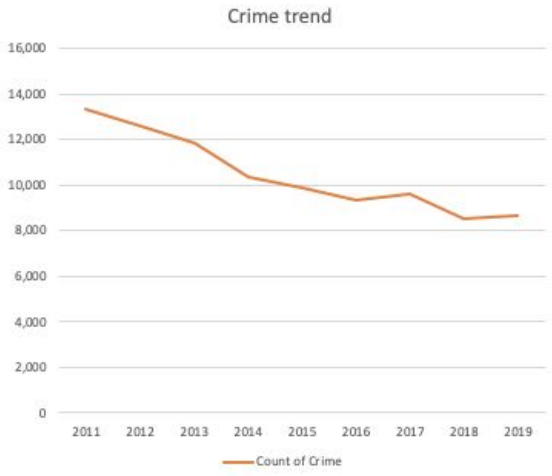


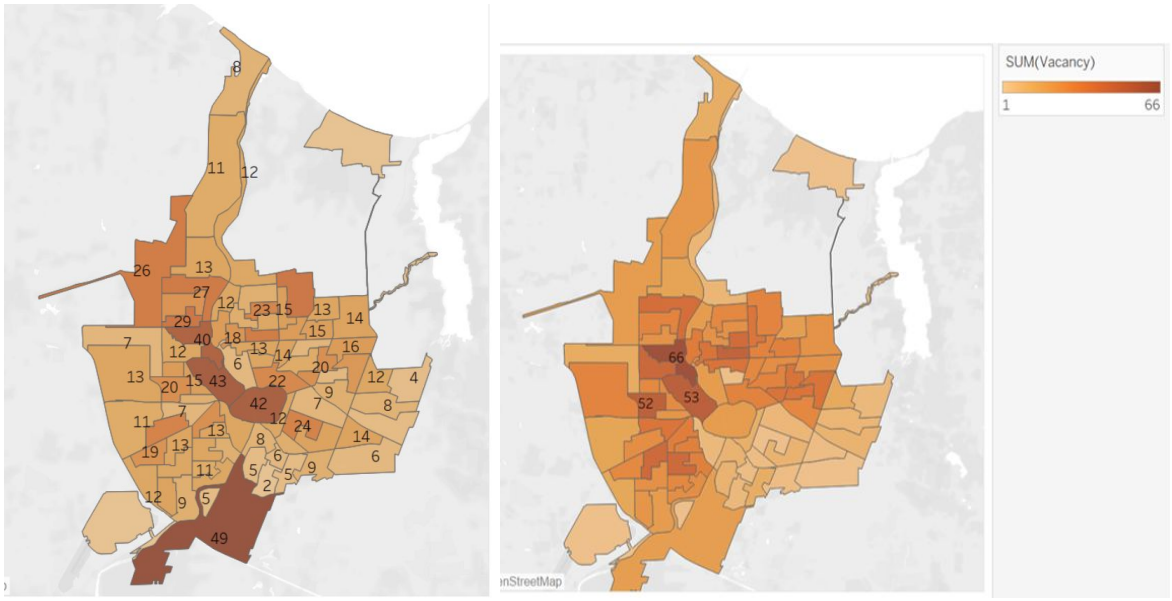
Fig. 2 Crime Trend 2011-2019

3.2 Correlation Analysis

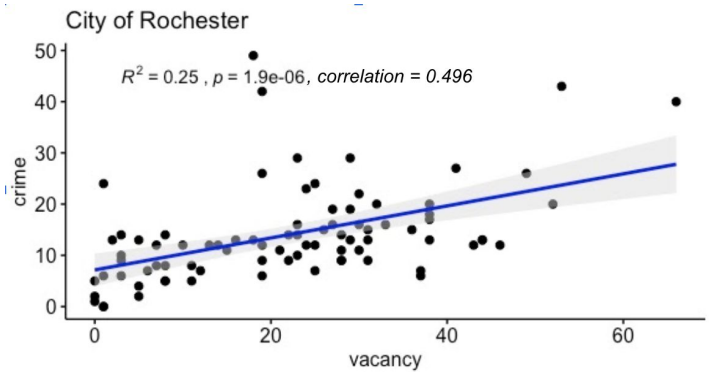
3.2.1 Crime and Vacancy

We aimed to find the correlation between crime and vacancy on tract level, so we selected the vacancy dataset and the crime dataset that contains past 60-day records to plot two choropleths respectively to represent the relation between these two features. As Fig.3 (a) and (b) indicate, there is a positive correlation between crime and vacancy as the color of two choropleths appeared to be in a similar pattern. Fig.3 (c) further confirms the positive correlation, while

showing that the correlation (correlation = 0.496, p = 1.9e-6) is not strong enough for us to build the model without further modification.



(a) Crime Choropleth (b) Vacancy Choropleth



(c) Correlation Scatterplot between Crime and Vacancy

Fig. 3 (a) the number of crime instances in past 60 days on tract level; (b) the number of currently existing vacant properties on tract level, and three tracts with the most number of vacancy are labeled; (c) the linear correlation between crime and vacancy

We then moved to explore whether there would be a more significant correlation between crime and vacancy by focusing on the change in the number of crime instances within one mile radius from the vacancy location one to six months before and after the vacancy appeared. On average, 51.6% vacancy instances had an increasing trend of crime within one mile radius over six months since the vacancy appeared. Thus, vacancy alone could not be used as a clear indicator of the surrounding crime.

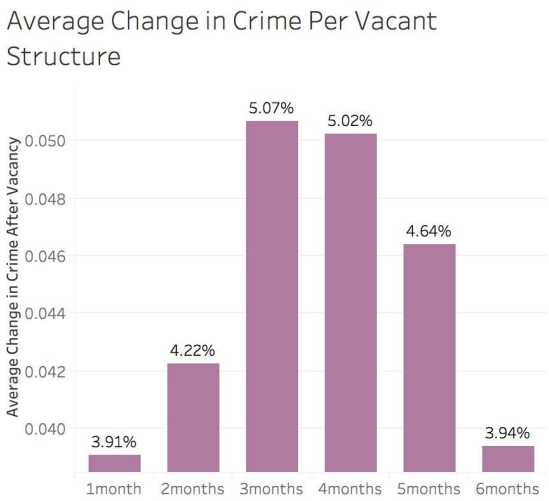


Fig. 4 Change in Crime within 1 Mile Radius from Vacancy Location

3.2.2 Crime and Demolition

Similarly, we explored the change in the number of crime instances within one mile radius from the demolition location one to six months before and after the demolition appeared. On average, 47.2% demolition instances had a decreasing trend of crime within one mile radius over six months since the demolition appeared. Thus, demolition alone could not be used as a clear indicator either of the surrounding crime.

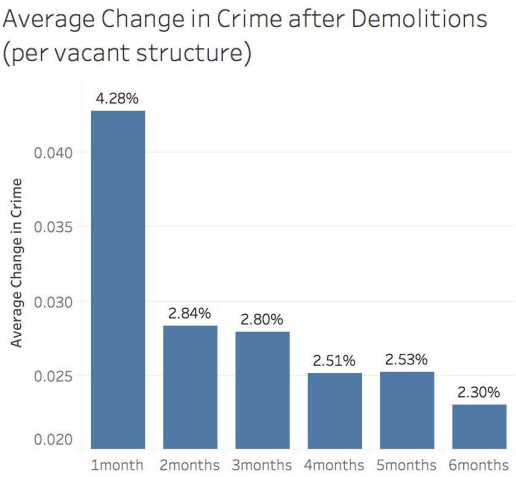
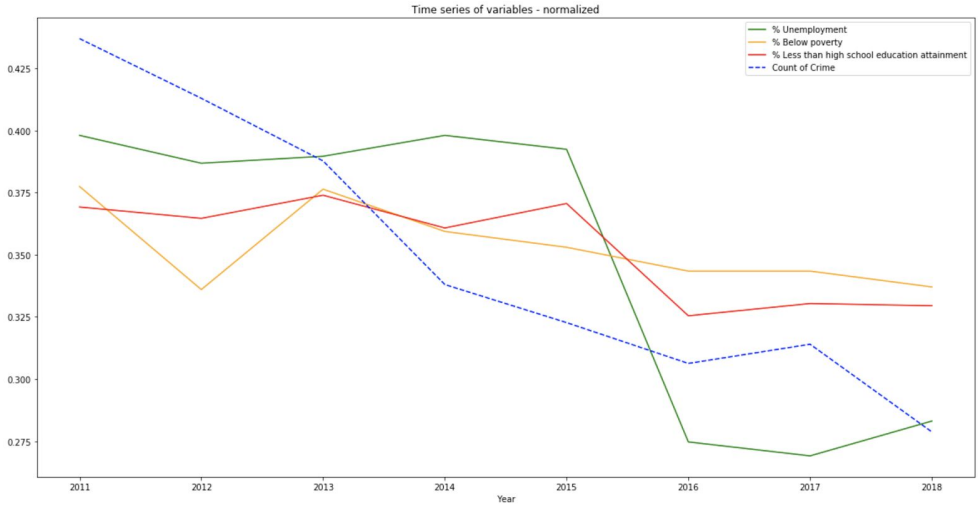


Fig. 5 Change in Crime within 1 Mile Radius from Demolition Location

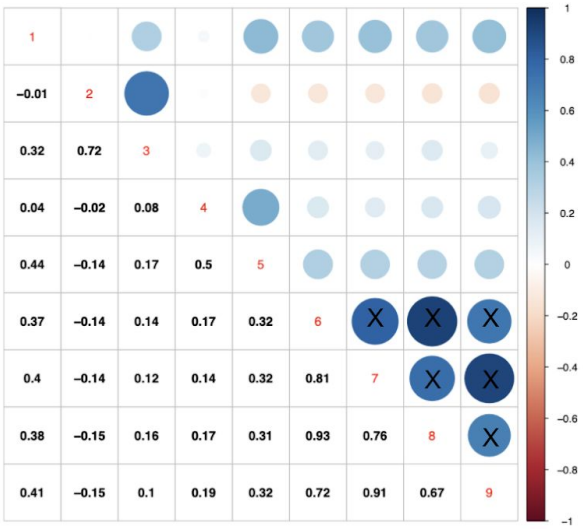
3.3 Integrated Census Features

Since neither vacancy nor demolition could serve as a clear indicator of crime rate, it was necessary to integrate other features from the census dataset.



(a) Trend of Selected Census Features and Crime

1.Number of Vacant Structures in the Tract
2.Population in the Tract (all age)
3.Education attainment in the tract total population below high school for >18 age
4.Unemployment status in the tract
5.Poverty Status in the tract
6.Crime ONE month BEFORE demolition
7.Crime ONE month AFTER demolition
8.Crime THREE months BEFORE demolition (accumulative)
9. Crime THREE months AFTER demolition (accumulative)



(b) Correlation among crime and selected census features

Fig. 6 (a) The trend of change of selected census features and crime from 2011 to 2019; (b) Correlation analysis for integrated features

Fig. 6 (a) showed the positive correlation among crime, unemployment rate, poverty rate, and education attainment. In Fig. 6 (b) the correlation among feature 6 to feature 9 were crossed out since all of them related to the crime before and after demolitions. Besides, Fig.6 (b) represented the positive correlation between feature 1 (number of vacant structures in the tract) and feature 5 to 9 (poverty status, crime one/three month(s) before and after demolition). Also, it represented a positive correlation between feature 5 (poverty status) and feature 6 to 9 (crime one/three month(s) before and after demolition). We also noticed that there was a negative correlation between feature 2 (population in the tract) and features other than feature 4 (Unemployment status). Such correlation could be caused by the increasing number of population in the city of Rochester and the decreasing trend for the other features.

4. MODEL DEVELOPMENT

As mentioned in the previous section, although there is a positive correlation between vacancy and the crime, the correlation is not very strong. We thought the vacancy alone cannot be a causing factor for crime, so we integrated other features such as education attainment, poverty status, unemployment and population into our model for predicting the number of crime after demolitions.

4.1 Modeling Setup

4.1.1 Data Forecasting

The American community surveys that we used included data from 2011 to 2018. As part of the new dataset, we needed to forecast the variables integrated from the census data source to reflect 2019 and 2020 numbers. General decreasing trend for each variable encouraged the use of extrapolated data instead of forecasting using constant 2018 numbers. For each variable's time series, we used the ARIMA model in predicting for each tract, after analysing the acf and pacf plots, and differencing to remove trends. Forecasted graphs followed the actual plots closely with low mean squared errors and thus the predictions were incorporated into the dataset for modelling.

4.1.2 Data Splitting

The dataset we built was split into a training dataset and a test dataset. 80% of the data was included in the training set and the rest of them were included in the test set.

4.1.3 Model Validation

K-fold cross-validation was used for model validation. The K was set to be 5. 5-fold cross-validation evaluated the model performance on different subsets of the training data and then calculated the average results.

4.1.4 Making Prediction

After we ran different models, the model with the highest performance was selected. We applied the model on the existing vacant structures dataset, and got predictive values of crime cases within a mile after 1 month of the demolitions. We calculated the rate of change in crime cases, and ranked the order for demolitions based on the rate of change. Then we created a feature called *order* to capture the recommended order for demolitions.

4.1.5 Evaluation Metrics

We used root mean squared error (RMSE) and R-squared value as our evaluation metrics. The detailed formula for RMSE and R-squared are:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

4.2 Methods

We used two linear models (linear regression and Lasso regression) and three non-linear models (Random Forest, Support Vector Machine and K-nearest Neighbor).

4.2.1 Linear Regression

The linear regression model was chosen to be our first prediction model. All of our variables were numerical and could fit in the linear regression model. The results of linear regression will be illustrated in section 5.2.

4.2.2 Lasso Regression

Lasso regression, or the Least Absolute Shrinkage and Selection Operator, is a modification of linear regression. In Lasso, the loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients (also called the L1-norm). The loss function is

$$\sum_{i=0}^{N} (y_i - \sum_{j=0}^{M} x_{ij} W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

where y_i is the dependent variable. x_{ij} are independent variables. W_j are parameters for corresponding independent variables. And λ is the penalty parameter for model coefficients. The optimal λ was selected using the *cv.glmnet* function in the *glmnet* package of R [14]. The results of Lasso regression will be illustrated in section 5.2.

4.2.3 Random Forest

Random Forest consists of multiple single decision trees. Each tree is based on a random sample of the training data. It is typically used to handle the overfitting problem and performs relatively well on the unforeseen dataset. The results of Random Forest will be illustrated in section 5.2.

4.2.4 Support Vector Machine (SVM)

SVM captures complex relationships between data records without having to perform difficult transformations manually. It is effective in high dimensional spaces. The results of SVM will be illustrated in section 5.

4.2.5 K-nearest Neighbor (KNN)

The classification of KNN is based on the voting of the majority of 'K' nearest instances. The distance used in KNN is usually Euclidean distance. The distance is affected by the scale of attributes. In order to eliminate the effect of the magnitude of attributes on the distance, we normalized the numerical attributes in the dataset. The normalization was realized using the *preProcess* function in the *caret* package of R [15]. The K was set to be 16, which was approximately equal to the square root of the number of samples in the training dataset. The results of KNN will be illustrated in section 5.

5. PERFORMANCE AND RESULTS

As shown in Table 1, the Random Forest model has the highest R-squared value, so we chose it to be the model for prediction.

Table 1. Model Performance

Test Models	RMSE	R-squared
Random Forest	16.982	0.620
Linear Regression	16.642	0.619
Lasso Regression	16.689	0.617
SVM	17.865	0.561
KNN (K=16)	18.671	0.521

We applied the Random Forest model on the existing vacant structures dataset, and got predictive values of crime cases within 1 mile after 1 month of the demolitions. Table 2 shows the results of the top 5 recommended vacant structures for demolitions based on the predictive rate of change in crime cases.

Table 2. Top 5 Recommended Vacant Structures for Demolitions

Address	predict_delta_perc	order
JEFFERSON AV 0560	-25.96%	1
N CLINTON AV1240	-25.14%	2
HUDSON AV 1070	-24.70%	3
HUDSON AV 1058	-24.45%	4
BURBANK ST 0003	-24.22%	5

6. CONCLUSION AND FUTURE WORKS

Our analysis shows that there is no strong correlation between the crime rate and number of vacant structures in the city of Rochester. Considering macroeconomic variables such as less than high school education attainment and poverty, gave way to building models, with Random Forest having the best performance. Based on this model, we establish demolition prioritizing which shows that the top recommendation predicts reduction in crime in the tract of the demolished vacant structure to be 25.96% as detailed in table 1 above.

Going forward, we hope to make the model more robust and tackle the challenges faced throughout the project. One of those being to gain access to the complete historical vacant structures data file along with the code book. This allows broader analysis on structures throughout the years and will show the trend over time. Integrating multiple datasets from various data sources meant that there was restriction in time frames of data. The US Census data provided annual data which could be modelled to reflect seasonality. In this project we considered a handful of variables and hope to incorporate more that have significant relationship with crime in each of the tracts.

From analysis on how vacant structures affect different crime categories, we found that from the 6 crime categories, aggravated assault (0.520) and motor vehicle theft (0.526) had the higher correlations, as abandoned properties could be used in these crimes. We could include a weighting on this factor in our model as well as changing the crime radius range from 1 mile to a smaller value like 0.2 mile when calculating before and after crime rates. This could better reflect the geography of an urban city and distance between properties.

7. REFERENCE

[1] "Preliminary Report." *FBI*, FBI, 20 Dec. 2019, ucr.fbi.gov/crime-in-the-u.s/2019/preliminary-report.

[2] "Rochester, NY Crime Rates." *NeighborhoodScout*, www.neighborhoodscout.com/ny/rochester/crime.

[3] "City of Rochester." *City of Rochester*, www.cityofrochester.gov/article.aspx?id=8589954382.

[4] Bothos, John M.A, and Stelios Thomopoulos. "Factors Influencing Crime Rates: an Econometric Analysis Approach." *Https://Www.researchgate.net/*, Researchgate, May 2016, www.researchgate.net/publication/303323965_Factors_influencing_crime_rates_an_econometric_analysis_approach.

[5] "Vacant Structures, Vacant Land, and City Owned Properties." *ArcGIS Web Application*, City of Rochester, maps.cityofrochester.gov/portal/apps/webappviewer/index.html?id=3c03bc009018476c93c564c013dfc484.

[6] "City of Rochester Demolition Tracker." *ArcGIS Web Application*, maps.cityofrochester.gov/portal/apps/webappviewer/index.html?id=ee6bd85f5cec43e098b9c824e448cc6e.
[7] "Rochester, NY Police Department Open Data Portal." *Data*, data-rpdny.opendata.arcgis.com/.
[8] "Education Attainment." *Data.census.gov*, data.census.gov/cedsci/table?q=Education Attainment &hidePreview=false&tid=ACSST1Y2018.S1501&t=Education&vintage=2018.
[9]"Poverty Status In The Past 12 Months Of Families." *Data.census.gov*, data.census.gov/cedsci/table?q=Poverty Status In The Past 12 Months Of Families&tid=ACSST1Y2018.S1701&t=Poverty.
[10]"American Community Survey (ACS) Why We Ask: Income." *Census Reporter: Making Census Data Easy to Use*, www2.census.gov/programs-surveys/acs/about/qbyqfact/2016/Income.pdf.
[11]"Selected Economics Characteristics." *Data.census.gov*, data.census.gov/cedsci/table?q=SELECTED ECONOMIC CHARACTERISTICS &tid=ACSDP1Y2017.DP03&vintage=2017.
[12]"American Community Survey (ACS) Why We Ask: Work Status." *Census Reporter: Making Census Data Easy to Use*, www2.census.gov/programs-surveys/acs/about/qbyqfact/2016/WorkStatus.pdf.
[13]"Acs Demographic and Housing Estimates." *Data.census.gov*, data.census.gov/cedsci/table?q=population&tid=ACSDP5Y2018.DP05.
[14] Friedman J, Hastie T, Tibshirani R. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, **33**(1), 1–22. 2016. www.jstatsoft.org/v33/i01/.
[15] Kuhn, Max, et al. *Classification and Regression Training*. cran.r-project.org/web/packages/caret/caret.pdf.