Homework 3

Group 1

Contents

Prepared for:

Dr. Nathan Bastian

City University of New York, School of Professional Studies - Data 621

Prepared by:

Group 1

Senthil Dhanapal

Yadu Chittampalli

Christophe Hunt

1 Introduction

Crime has a high cost to all parts of society and it can have severe long term impact on neighborhoods. If crime rises in the neighborhood or it is invaded by criminals, then families and those with the economic means to leave for more stable areas will do so[1]. Additionally, crime can even have a health cost to the community in that the perception of a dangerous neighborhood was associated with significantly lower odds of having high physical activity among both men and women[2]. It is important to understand the propensity for crime levels of a neighborhood before investing in that neighborhood.

2 Statement of the Problem

The purpose of this report is to develop a statistical model to determine the variables that are independently associated with neighborhoods with crime rates above or below the median. Note that neighborhoods with crime rates above or below the median have been provided in our evaluation data set.

3 Data Exploration

3.1 Variables Explained

The variables provided in our evaluation data set our explained below:

Abbreviation	Definition
zn	proportion of residential land zoned for large lots (over 25000 square feet)
indus	proportion of non-retail business acres per suburb
chas	a dummy var. for whether the suburb borders the Charles River (1) or not (0)
nox	nitrogen oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment centers
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
black	1000(Bk - 0.63)2 where Bk is the proportion of blacks by town
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000s

[1] Effect of Crime on Real Estate Values. (1952). The Journal of Criminal Law, Criminology, and Police Science, 43(3), 357-357. Retrieved from http://www.jstor.org.remote.baruch.cuny.edu/stable/1139159

[2] Bennett GG, McNeill LH, Wolin KY, Duncan DT, Puleo E, Emmons KM (2007) Safe To Walk? Neighborhood Safety and Physical Activity Among Public Housing Residents. PLoS Med 4(10): e306. doi:10.1371/journal.pmed.0040306

3.2 Exploration of Variables

The skewness of each input variable is shown below. The two variables with the strongest skew are the proportion of residential land zoned for large lots and the proportion of blacks by town. Respectively the magnitudes of the skewness of these two variables are 2.18 and 2.92. This indicates that the distributions for these two variables are far from symmetrical. The skewness of the dummy variable (whether the suburb borders the river or not) can be neglected because it is a binary variable. All of the other variables skewnesses that are approximately of magnitude 1 or less. This indicates that the distributions for those variables can be considered symmetric even though for three of the variables (concentration of nitrogen oxides, index of accessibility to radial highways, and median value of owner-occupied homes) are multimodal.

variables	skew
zn	2.1768152
indus	0.2885450
chas	3.3354899
nox	0.7463281
rm	0.4793202
age	0.5777075
dis	0.9988926
rad	1.0102788
tax	0.6593136
ptratio	0.7542681
black	2.9163108
lstat	0.9055864
medv	1.0766920
target	0.0342293

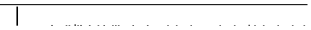
According to the standard deviations of each variable, the variable that has the highest difference from the mean is tax.

variables	sd
zn	23.3646511279634
indus	6.84585491881262
chas	0.256791996193711
nox	0.116666665669521
rm	0.704851288243787
age	28.3213784029166
dis	2.10694955535994
rad	8.68592724130043
tax	167.900088684704
ptratio	2.19684473073614
black	91.3211298387792
lstat	7.10189067779907
medv	9.23968141143397
target	0.500463581298941

Histograms of most of our variables have been plotted below so that distribution can be visualized. We have excluded `target` and `chas` due to being binary and not being well represented in the below visualization. We also excluded `rad` as it is an index variable and also is not best represented in the below visualization.

Table 1 : Descriptive Statistics
11 Variables 466 Observations

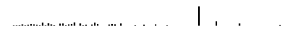
zn



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	26	0.61	12	0	0	0	0	16	45	80

`lowest :   0  12  18  18  20, highest:  82  85  90  95 100`

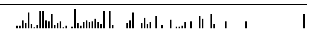
indus



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	73	0.98	11	2	3	5	10	18	20	21

`lowest :  0.5  0.7  1.2  1.2  1.2, highest: 18.1 19.6 21.9 25.6 27.7`

nox



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	79	1	0.6	0.4	0.4	0.4	0.5	0.6	0.7	0.8

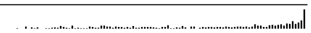
`lowest : 0.4 0.4 0.4 0.4 0.4, highest: 0.7 0.7 0.7 0.8 0.9`

rm

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	419	1	6	5	6	6	6	7	7	8

`lowest : 4 4 4 5 5, highest: 8 8 9 9 9`

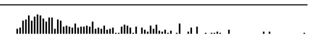
age



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	333	1	68	18	26	44	77	94	99	100

`lowest :   3   6   6   6   7, highest:  99  99  99  99 100`

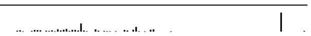
dis



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	380	1	4	1	2	2	3	5	7	8

`lowest :  1  1  1  1  1, highest:  9  9 11 11 12`

tax



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	63	0.98	410	222	233	281	334	666	666	666

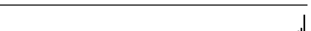
`lowest : 187 188 193 198 216, highest: 432 437 469 666 711`

ptratio

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	46	0.98	18	15	15	17	19	20	21	21

`lowest : 13 13 14 14 15, highest: 21 21 21 21 22`

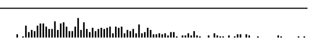
black



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	331	0.99	357	88	295	376	391	396	397	397

`lowest :   0.3   2.5   2.6   3.5   3.6`
`highest: 396.3 396.3 396.3 396.4 396.9`

lstat



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	424	1	13	4	5	7	11	17	23	27

`lowest :  2  2  2  2  3, highest: 34 34 35 37 38`

medv

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
466	0	218	1	23	10	13	17	21	25	35	43

`lowest :  5  6  6  7  7, highest: 46 47 48 49 50`

3.3 Correlation Matrix

We implement a correlation matrix to better understand the correlation between variables in the data set. The below matrix is the results and we noticed a few interesting correlations.

- High nitrogen oxides concentration (parts per 10 million) ("nox") is positively correlated with higher than median crime rates. As defined by the EPA - "NOx pollution is emitted by automobiles, trucks and various non-road vehicles (e.g., construction equipment, boats, etc.) as well as industrial sources such as power plants, industrial boilers, cement kilns, and turbines"[3]. It is clear to see that nox is concentrated in areas of high road traffic and possible high industrial use which would be neighborhoods of low value and may attract crime.
- The weighted mean of distances to five Boston employment centers is negatively correlated with a city with higher than median crime rate. This is intuitive in that employment centers would be more closely located in cities of high crime due to high unemployment being positively correlated with higher crimes rates[4].
- The tax is positively correlated with higher than median crime rate which is counter intuitive because we would think as tax increases then crime would decrease (more valuable property = higher tax = less crime).
- We also see bk is negatively correlated with higher than median crime rates but it seems to be due to the transformation of 1000(Bk - 0.63)ˆ2. Further resources on why this type of transformation is being used were not available. It should be noted that this transformation causes a counter intuitive correlation.

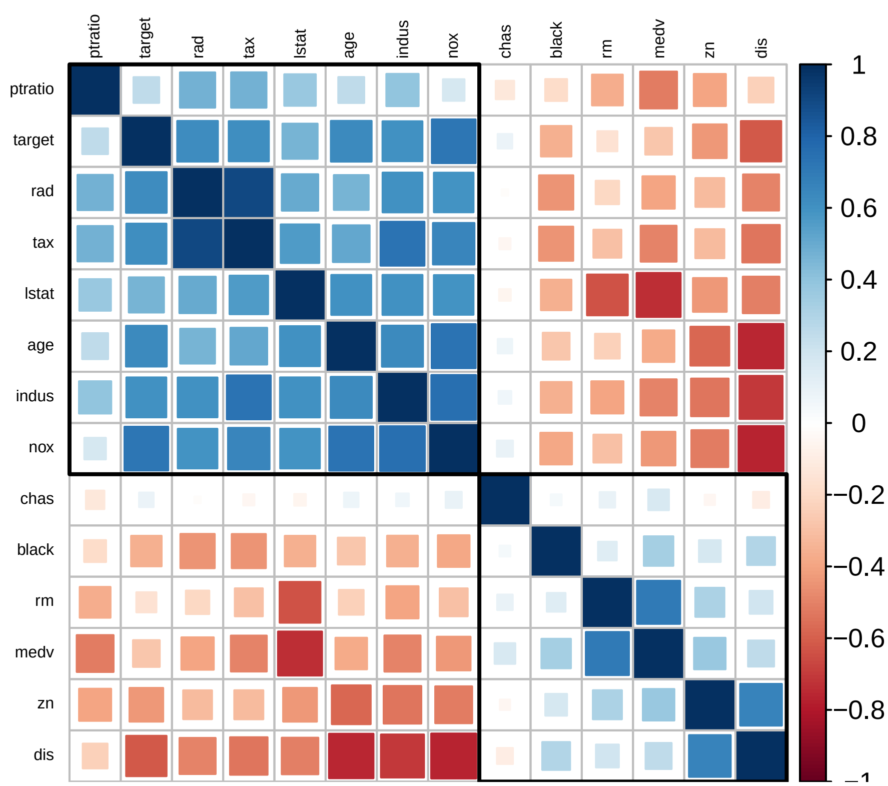


Figure 1: Correlation Plot of Training Data Set

[3]"Nitrogen Oxides Control Regulations | Ground-level Ozone | New England | US EPA." EPA. Environmental Protection Agency, n.d. Web. 22 Oct. 2016.

[4]Ajimotokin, S., Haskins, A., & Wade, Z. (2015). The Effects of Unemployment on Crime Rates in the US.

3.4 Outliers Treatment

We chose winsorizing as the method to address outliers. Instead of trimming values, winsorizing uses the interquantile range to replace values that are above or below the interquantile range multiplied by a factor. Those values above or below the range multiplied by the factor are then replaced with max and min value of the interquantile range. Using the factor 2.2 for winsorizing outliers is a method developed my Hoaglin and Iglewicz and published Journal of American Statistical Association in 1987[5].

The below table is the summary results of the winsorizing of the data.

Table 4:

Statistic	N	Mean	St. Dev.	Min	Max
zn	466	−4.024	5.144	−7.162	5.024
indus	466	1.942	0.662	−0.804	2.882
chas	466	0.071	0.257	0	1
nox	466	−0.879	0.367	−1.571	−0.148
rm	466	1.900	0.117	1.518	2.201
age	466	2,735.940	1,734.456	3.705	4,997.945
dis	466	0.802	0.269	0.117	1.253
rad	466	1.324	0.464	0.000	1.955
tax	466	0.997	0.001	0.995	0.999
ptratio	466	171.129	38.664	78.868	241.452
black	466	72,528.400	8,276.700	54,853.990	78,731.590
lstat	466	1.910	0.401	0.522	2.672
medv	466	3.574	0.541	1.753	4.582
target	466	0.491	0.500	0	1

[5]Hoaglin, D. C., and Iglewicz, B. (1987), Fine tuning some resistant rules for outlier labeling, Journal of American Statistical Association, 82, 1147-1149.

4 Models Built

4.1 Model 1 - Backwards Selection Method

Table 5:

	Dependent variable:
	fullModel
nox	0.742***
	(0.088)
age	0.0001***
	(0.00001)
dis	0.264**
	(0.112)
rad	0.221***
	(0.044)
tax	66.782***
	(21.918)
black	−0.00001***
	(0.00000)
medv	0.137***
	(0.033)
Constant	−66.185***
	(21.904)
Observations	466
Log Likelihood	−97.263
Akaike Inf. Crit.	210.527
Note:	*p<0.1; **p<0.05; ***p<0.01

4.2 Model 2 - Forwards Selection Method

4.3 Model 3 - Subset Selection Method

Using the `leaps` package and the `regsubsets` function we are able to subset our independent variables by looking at the best model for each predictor. Our final model using this method is line `8` which we will further implement into our subset selection model.

	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1 (1)				*									
2 (1)				*				*					
3 (1)				*		*		*					
4 (1)				*		*		*					*
5 (1)				*		*		*		*			*
6 (1)				*		*		*		*	*		*
7 (1)				*		*		*	*	*	*		*
8 (1)				*		*		*	*	*	*	*	*

Table 7:

	Dependent variable:
	target
nox	33.412***
	(5.244)
age	0.019*
	(0.010)
rad	0.734***
	(0.142)
tax	$-$0.010***
	(0.003)
ptratio	0.364***
	(0.110)
black	$-$0.012*
	(0.007)
lstat	0.058
	(0.047)
medv	0.094***
	(0.036)
Constant	$-$24.903***
	(5.027)
Observations	466
Log Likelihood	$-$101.509
Akaike Inf. Crit.	221.018
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01