Case Study 3 (CS3): Linear Crime Models

"On my honor, I pledge that I have neither given nor received help on this assignment."

Name: Prateek Agrawal
Computing ID: pa7sb
Date: 10th October 2015
Course: Data Mining(SYS-6018)

Situation

The kernel density estimate (KDE) approach to crime analysis is nice in that it is simple to implement and easily interpretable. However, the Chicago Police Department (CPD) is unsatisfied with the KDE approach because it does not tell them anything about why crime patterns look like they do. They are asking the following questions:

Question 1.

Intuitively, what spatial factors might influence the occurrence of theft or assault (you can pick one)? Why? You should include at least one reference to the research literature in support of your answer, and you should discuss at least three factors?

Solution:

For this problem we have chosen the crime case which involves 'Theft' reports only.

There can be many spatial factors that can influence the occurrence of 'Theft', the most important among them would be

a.) The Distance from schools
b.) The location of hospitals and
c.) The location of Police Stations.

We will discuss each of these factors

1. ***The Distance from Schools,***

The distance from the closest school have a significant impact on the crime rate, especially on school days. This can be interpreted as majority of the youth will be at or near school during the school day, so there is a high probability that the crime concentration will be higher in blocks situated close to the schools. Also the likelihood of the interactions and hangout between violent individual increases, as we get near to a school. There are many research articles and books relating to this hypothesis like;

1. 'Schools, Neighborhoods, ad Violence: *Crime within the Daily Routines of Youth*' by 'Caterina Gouvis Roman',

https://books.google.com/books?id=ejcZK8aHWg0C&pg=PA69&lpg=PA69&dq=research+article+on+crime+vs+distance+from+school&source=bl&ots=5rdLLnhWd2&sig=ENjEgk2E9soYlbFUgjzmPpKdcr4&hl=en&sa=X&ved=0CDwQ6AEwBGoVChMI5-6RiLK7yAIVg6w-Ch27HArv#v=onepage&q=research%20article%20on%20crime%20vs%20distance%20from%20school&f=false

2. A research paper on 'HIGH SCHOOLS AND CRIME: A REPLICATION' by Dennis W. Roncek and Donald Faggian.

http://onlinelibrary.wiley.com/doi/10.1111/j.1533-8525.1985.tb00240.x/abstract

2. ***The Distance from Hospitals,***

The distance from the hospital is also inversely proportional i.e as the minimum distance increases the likelihood of theft decreases. As can be seen as there is always a higher concentration of people present in and around hospitals, therefore, there will be more likelihood of crime in blocks situated near the hospitals.

a.) Street Robbery: Khadija M. Monk, Justin A. Heinonen, John E. Eck

http://www.popcenter.org/problems/pdfs/street_robbery.pdf

b.) Health care crime survey by IHSS foundation http://ihssf.org/PDF/crimesurvey2014.pdf

3. ***The Distance from Police Stations,***

The obvious guess would be that as the minimum distance increases the theft reports should increase, but police stations are strategically placed in high crime concentration zone, therefore, the theft

reports would be inversely proportional to the minimum distance of police station, i.e. as the distance increases the theft reports decreases. There are many research articles relating that supports this hypothesis

c.) Effectiveness of Police in Reducing Crime and the Role of Crime Analysis
https://us.sagepub.com/sites/default/files/upm-binaries/46974_CH_3.pdf

---

Question 2.

Use linear models to quantify the importance of the factors you discussed in your answer to question 1?

Approach:

1. I have used Logistic regression to perform linear model analysis of the crime data with theft density.
2. Minimum distance from a school, minimum distance from a hospital and minimum distance from a police station as regressors/variables/factors.
3. We also check for the multicollinearity among the regressors which could affect the response.
4. the significance of each regressors can be confirmed based on the p-values obtained from the model
5. To quantify the importance of each regressor/factor, I then used the model to predict response and then plot the predicted crime locations using a Surveillance plot, The values of Area under the curve gives the accuracy of the model.

Data:

The *theft crime data* is obtained from the Chicago Data portal (https://data.cityofchicago.org/), The data contains the reported cases of crime from 01/01/2014 12:00:00am to 12/31/2014 11:58:00pm. It has 22 different fields ranging from the crime ID, Case Number, ICUR code, Time, Location, Longitude, Latitude, Description of Crime etc. The data contains 56885 reported crime cases.

The *shapefiles* for schools, hospitals and police station can also be obtained from Chicago Data Portal (https://data.cityofchicago.org/browse?q=shapefile&sortBy=relevance&utf8=%E2%9C%93). They provide exact location of school, hospital or police station in the Chicago map.

Analysis:

1. I created 3 subsets of the theft crime data on the basis of theft reports per months. First is the theft reports from the month of Januray to March, Second is the theft reports from the month of April and the third is the theft reports for the month of May.
2. The model is trained on the theft reports obtained from the months of *January to March*. While the model reponses are fitted from the theft reports in the month of *April*, and finally the model is then used to predict the theft reports for the month of '*May*'.
3. The distance from closest school, the distance from closest hospital and the distance from closest Police station along with theft density are used as regressors/variables.
4. The significance of the each regressor can be checked with the p-values obtained from the model.
5. The model is also checked for multicollinearity, whether if there exists some relation between the variables which could affect the variance and response.
6. The low VIF values found in the model and low p-values (< 0.05) highlights, that all regressors are significant and independent. As shown in the figure

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.087e+00  4.956e-02 -21.923  < 2e-16 ***
theft.density                2.425e+08  8.697e+06  27.879  < 2e-16 ***
school.min.distance         -4.552e-04  5.567e-05  -8.176 2.92e-16 ***
hospital.min.distance       -1.170e-04  1.449e-05  -8.079 6.55e-16 ***
policestation.min.distance -3.451e-05  1.681e-05  -2.053     0.04 *
---
```

Figure1: showing the p-value for all the factors in the 1st model

```
      theft.density       school.min.distance      hospital.min.distance
           1.130722                  1.195573                  1.347874
policestation.min.distance
           1.366351
```

Figure2: the VIF values for the factors/regressors

Evaluation:

1. The model obtained from the analysis is
   Thefts = - 1.087 + 2.425e+08 * Theft density - 4.552e-04*(School.min.distance) - 1.170e-04 *(Hospital.min.distance) -3.451e-05* (Policestation.min.distance)
   The following things can be observed
   a.) *All the three factors minimum distance from school, minimum distance from hospital and minimum distance from police stations are inversely proportional to theft concentration, i.e. as the distance increases the theft decreases.*
   b.) *Among the three, 'School' have the major significance on the theft concentration.*
2. For evaluation of our model we used a surveillance plots which uses ROC curves that plots the true positive rate to false positive rate. Theoretically it can be written as:
   FPR = False Positive/Ground Negative
   TPR= True Positive/Ground Positive;
3. The Area under the ROC curve is used to quantify the importance of the regressors.

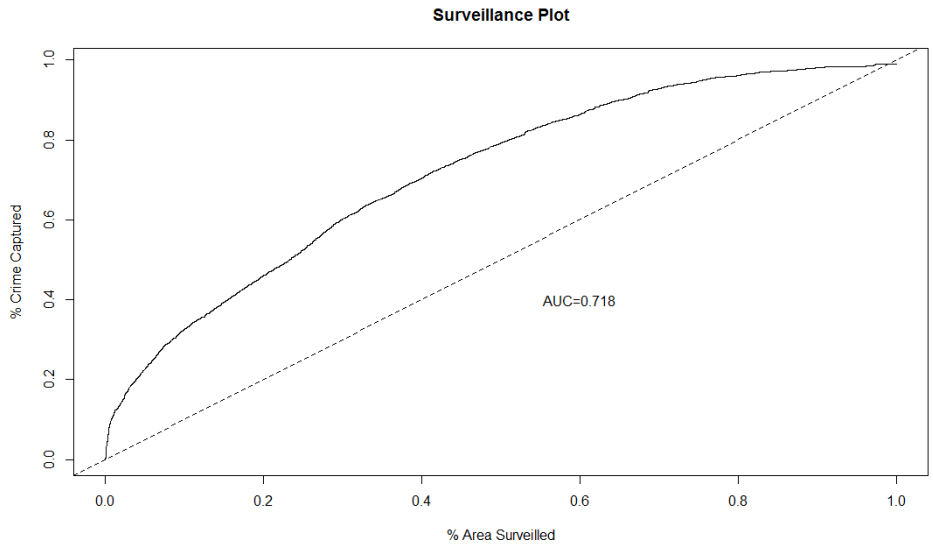

Figure3: The Surveillance curve for the full model.

4. Then I have created three different models i.e. a model for each regressor, to plot the surveillance plots to further confirm the significance and importance that each factor/regressor holds.

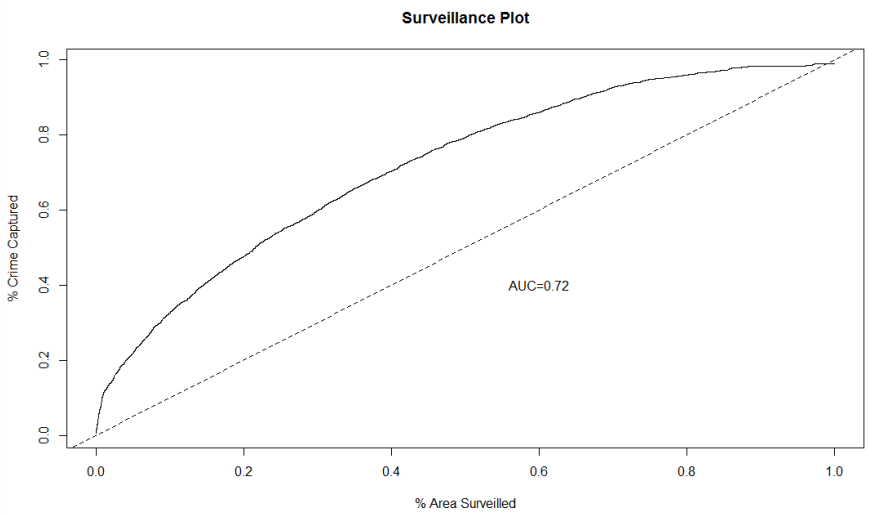

Figure4: The Surveillance plot for School.min.distance regressor

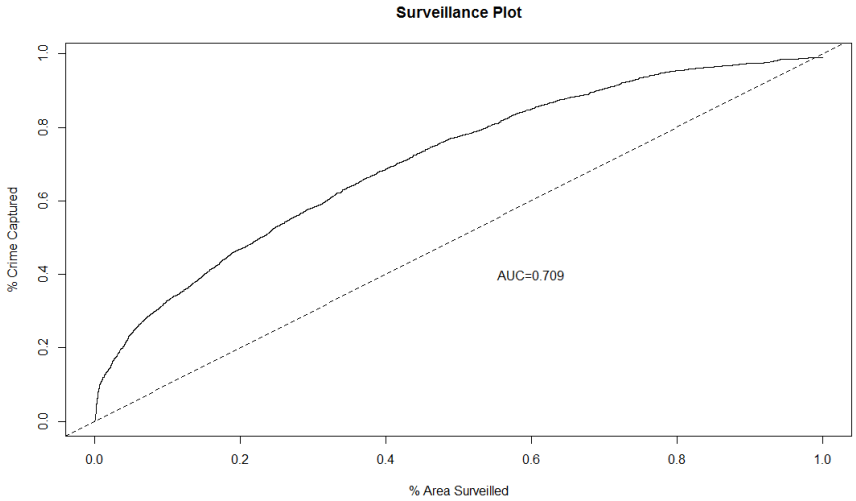

Figure5: The Surveillance plot for Hospital.min.distance regressor

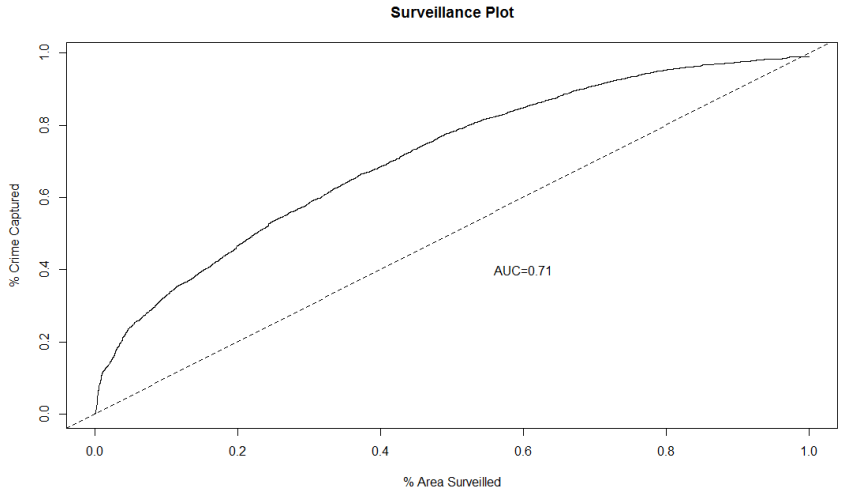

Figure6: The Surveillance plot for Policestation.min.distance factor

From the AUC curves we can observe that

a.) The high AUC values of each of these plots confirm that each of them is an importance factor.

b.) The minimum distance from school have the maximum value of the AUC, which confirms the previous observation that among the three factors, proximity to school holds the major importance to theft related crimes.

Question 3.

How should the CPD adjust their patrolling strategies to account for the factors you have identified?

Solution:

Recommendations:

1. The schools (especially the high schools) should be patrolled on school days and more significance should be laid on after school hours in the day time, there isn't much patrolling need in the night time near schools as most of the students will not be present at that time.
2. There should be constant patrolling in the blocks near the hospitals, as there is high likelihood of theft occurring in such locations.
3. As per our intuition that police station are located at high crime concentration area, therefore, patrolling should be heavy in areas near to the police station, and after maintaining a hold at the crime rate in those blocks, the police patrolling should expand.

****