# Case Study 4 (CS4): Non-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*
*Course: Data-Mining (sys-6018)*

# *Situation*

The Chicago Police Department (CPD) is not satisfied with the performance of your linear crime models, as measured by the surveillance plots. Their analysts point out that logistic regression is one of many available classifiers, and they ask for a comparison with other methods. Specifically, they are interested in comparisons with non-linear classifiers. You agree to evaluate the performance of non-linear support vector machines and compare these models with your logistic regression models.

1. *Hypothesize at least one predictor that might have a non-linear effect on the occurrence of crime. Locate data for this factor in the Chicago Data Portal. If you cannot locate such data, you may substitute an alternative factor.*

*Solution:*

I figure that the location of hospitals have a non-linear effect on the occurrence of crime because there are very few hospitals in Chicago as compared to number of theft locations in the city. Therefore, there is no way to definitively suggest that hospitals in the city of Chicago can have any linear effect on the locations of theft. The above hypothesis can also be suggest by the figure below..

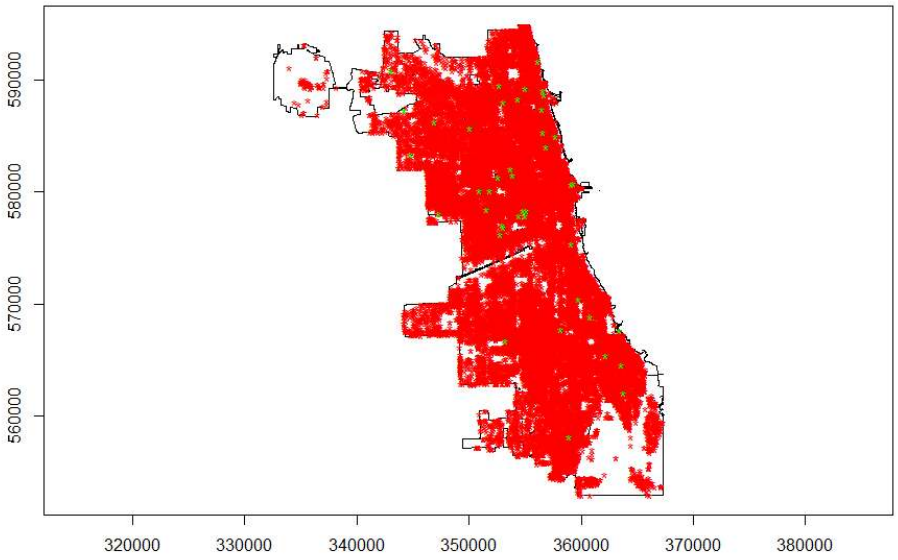


Figure1: The red points show the location of theft cases reported in the year 2014 and green points represent the location of hospitals in Chicago.

---

2. *Using the factor from (1) plus two additional factors (e.g., those used in CS3), build the following SVM models and compare their performance with logistic regression:*

a) *A linear SVM*
b) *A degree-2 polynomial kernel SVM*
c) *An RBF kernel SVM*

*Solution:*

For this problem I have chosen the crime case which involves 'theft' cases only. There can be many spatial factors that can influence the occurrence of 'Thefts' in the city, notably some that I have used in the following analysis are:

a) Hospitals
b) Police Stations
c) Railway Stations

***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 hospital, Police Stations and Railway Stations can also be obtained from Chicago Data Portal (https://data.cityofchicago.org/browse?q=shapefile&sortBy=relevance&utf8=%E2%9C%93).

For the purpose of analysis, we have used the theft reports from the month of January to Mach as training data, the model is fitted using the data from the month of April's theft reports and then predictions are made for the month of May's.

***I have performed both the Logistic regression and SVM using 2 different sets of regressors.***

***In the 1st analysis Spatial factors along with Kde is used to predict the thefts. While in the 2nd analysis only Spatial factors are taken into account.***

## *1st Analysis*:-

### _____*The Logistic Model*_____

***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 multi-collinearity 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

Surveillance plot. The value of Area under the curve gives the accuracy of the model.

***Analysis:***

1. The model is trained on the theft reports obtained from the months of April and the old predictions is used from the months of *January to March* and finally the model is then used to predict the theft reports for the month of '*May*'.
2. The distance of crime from the closest hospital, police station and railway stations and theft density are used as regressors/variables.
3. The significance of the each regressor can be checked with the p-values obtained from the model.

```
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.249e+00  5.327e-02 -23.439  < 2e-16 ***
theft.density                2.947e+08  1.127e+07  26.135  < 2e-16 ***
hospital.min.distance       -9.662e-05  1.514e-05  -6.384 1.73e-10 ***
policestation.min.distance  -6.952e-05  1.420e-05  -4.895 9.85e-07 ***
railwaystation.min.distance -6.839e-05  1.201e-05  -5.695 1.23e-08 ***
---
```

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

4. The model is also checked for multicollinearity, whether there exists some relation between the variables which could affect the variance and response.
5. 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

```
          theft.density       hospital.min.distance  policestation.min.distance
               1.359792                    1.573404                    1.440716
railwaystation.min.distance
               1.406881
```

Figure2: the VIF values for the factors/regressors

***Evaluation:***

1. The model obtained from the analysis is
   ***Response = - 1.249 + 2.947e+08 * Theft density – 9.662e-05*(hospital.min.distance) – 6.952e-05 *(policestation.min.distance) -6.839e-05* (railwaystation.min.distance)***
   The following things can be observed
   1 According to the model created all the three factors are inversely proportional to the response.
   2 The minimum distance from the hospital has the maximum effect on the response

   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;

2. The Area under the ROC curve is used to quantify the importance of the regressors.

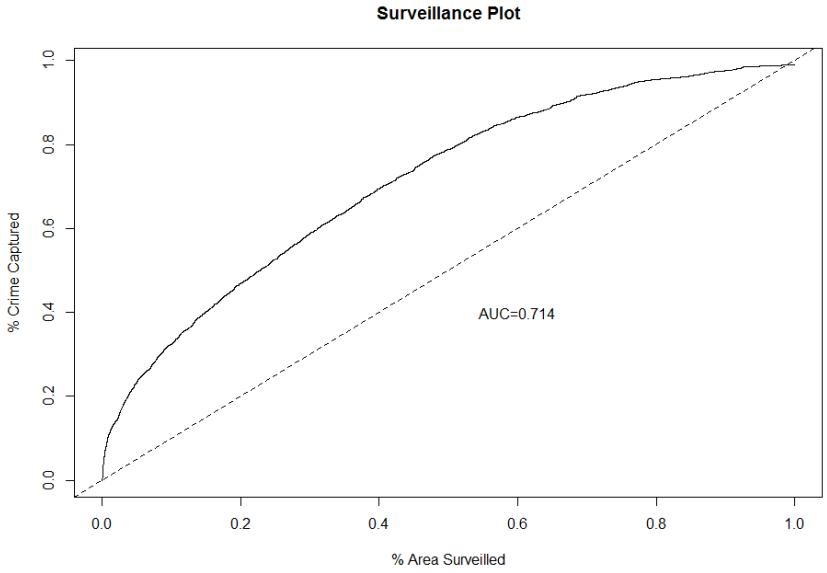


Figure 2: The Surveillance curve for the model.

From the AUC curves we can observe that

a.) The high AUC values of each of this plot confirm that the model explains the response i.e. the theft occurring in the city of Chicago.

## _____*The Support Vector Model*_____

***Approach:***

1 Here I have used SVM (Support Vector Machines) for the classification of the response as positive (crime occurring) or negative (no crime).

2 The minimum distance of a crime location from hospital, Police station and railway station along with theft density is used as regressors.

3 The response is then trained from the crime reports for the month of April.

4 Three SVM models are fitted. Namely

   a) The first is the Linear SVM.
   b) The second model is the degree-2 polynomial kernel SVM.
   c) Last is the RBF kernel SVM.

5 Some attributes regarding each of the models are.

   a) The default C-svc is used for classification
   b) The kernel function used in training and predicting is set according to the model i.e. 'Vanilladot', 'Polydot' & 'rbfdot'.
   c) The cost of constraints violation is set to 10 (due to computing power constraints, can be increased to 50 or 100).

6 To quantify the importance of each regressor/factor, I then used the model to predict response and then plot the predicted crime locations using

Surveillance plot. The value of Area under the curve gives the accuracy of the model.

***Analysis:***

1 The model is trained on the theft reports obtained from the months of *January to March*. While the model responses 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*'.

2 The distance of crime from the closest hospital, police station and railway stations and theft density are used as regressors/variables.

***Evaluation:***

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.

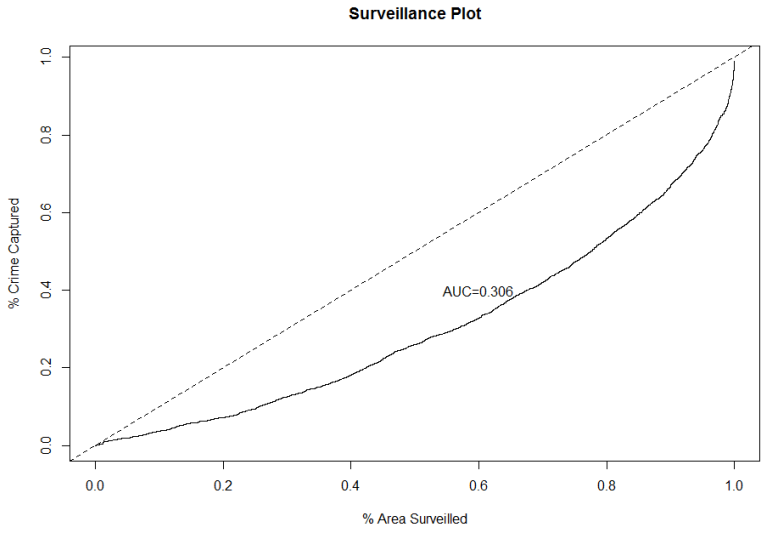


Figure 6: The Surveillance curve for linear SVM

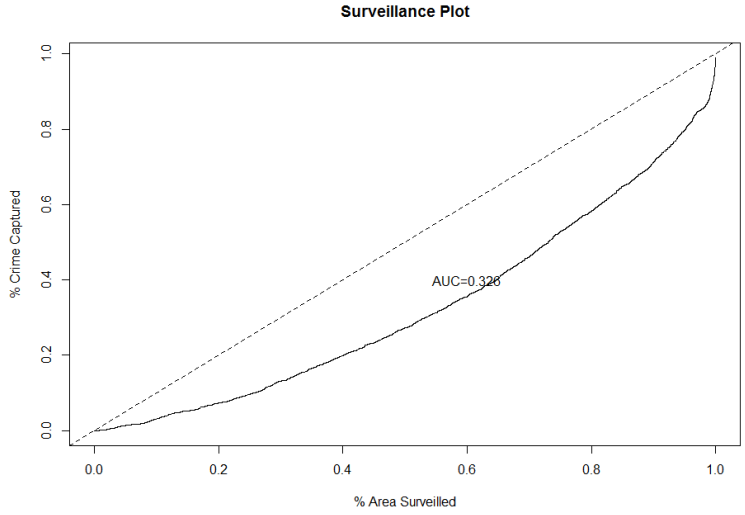


Figure 4: Surveillance plot for 2-degree polynomial kernel SVM

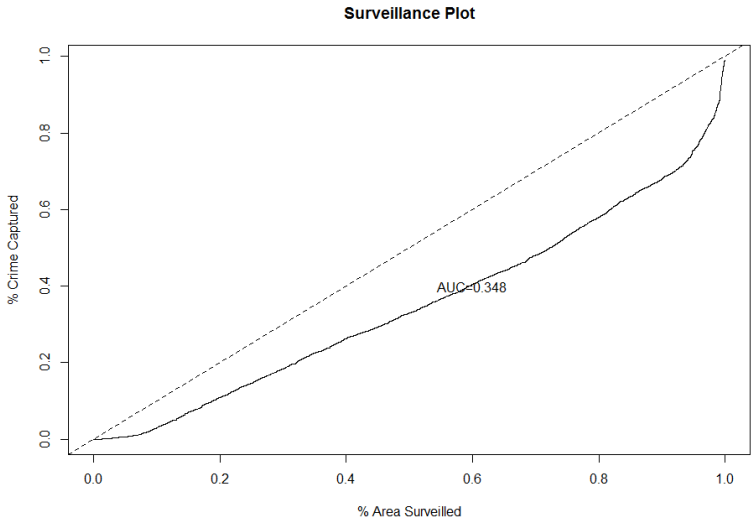


Figure 5: Surveillance plot for RBF kernel SVM

The AUC values are as follows:

1. Logistic regression: 0.714
2. The linear SVM: 0.306
3. The 2-degree polynomial kernel SVM: 0.326
4. RBF kernel SVM: 0.348

---

## *2nd Analysis*:

### ***The Logistic Model***

The approach and analysis format remains the same as the previous analysis method, only the type and number of regressors have been changed, here I have used only the spatial factors as regressors namely.

1. Minimum distance from closest hospitals,
2. Minimum distance from closest Police Stations
3. Minimum distance from Closest railway station

The model comes out to be:

***response = -1.446e-01 - 1.765e-04*(hospital.min.distance) – 8.675e-05*(policestation.min.distance) -1.990e-04*(railwaystation.min.distance)***

With p-values to be:

```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.446e-01  3.462e-02  -4.177 2.96e-05 ***
hospital.min.distance      -1.765e-04  1.465e-05 -12.051  < 2e-16 ***
policestation.min.distance -8.675e-05  1.376e-05  -6.303 2.91e-10 ***
railwaystation.min.distance -1.990e-04  1.153e-05 -17.262  < 2e-16 ***
---
```

The model is also checked for multi collinearity, with VIF values:

```
   hospital.min.distance  policestation.min.distance railwaystation.min.distance
                1.406213                    1.365322                    1.147036
```

Plotting the Surveillance plot for the model:

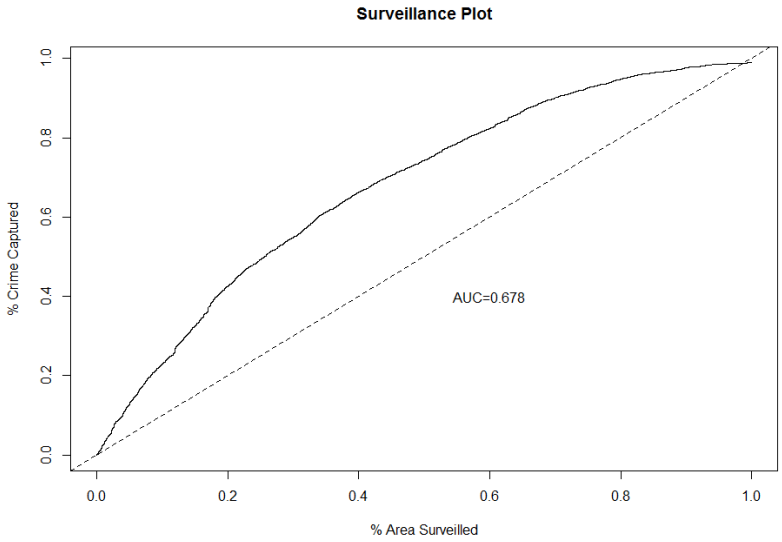


Figure 2: The Surveillance curve for the model.

## _____*The Support Vector Model*_____

While performing this model also, the kde regressor is not used and only the Spatial factors is used. The approach and analysis still remains the same as for 1st analysis.

Plotting the Surveillance plots for 3 different SVM methods.

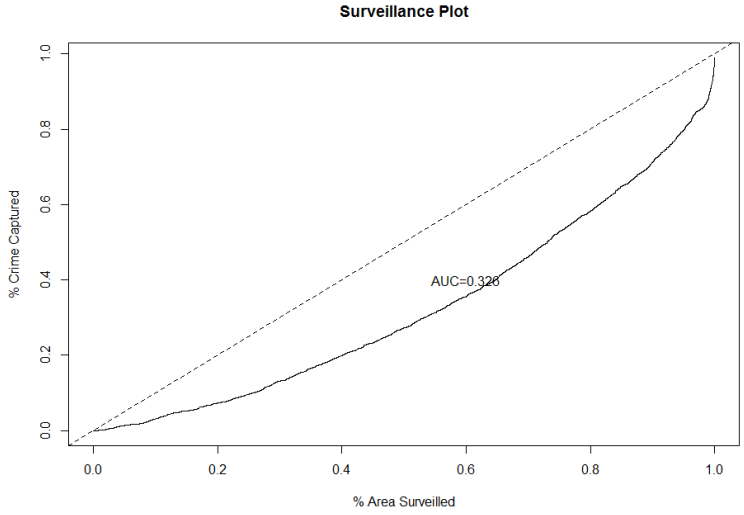


Figure 6: The Surveillance curve for linear SVM

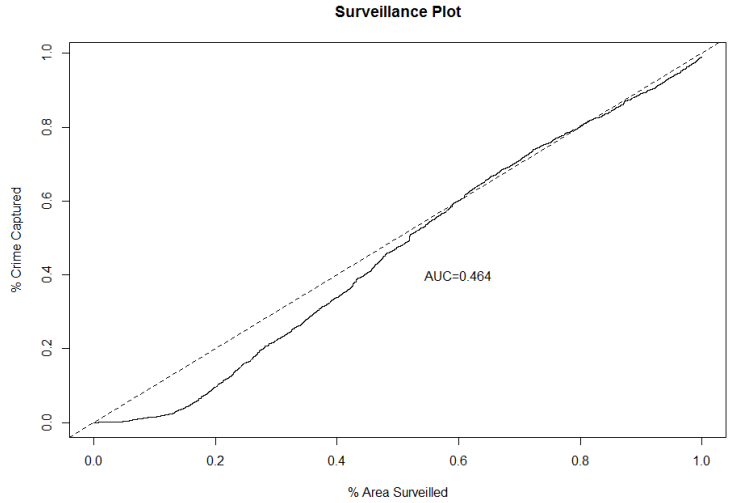


Figure 7: Surveillance plot for 2-degree polynomial kernel SVM

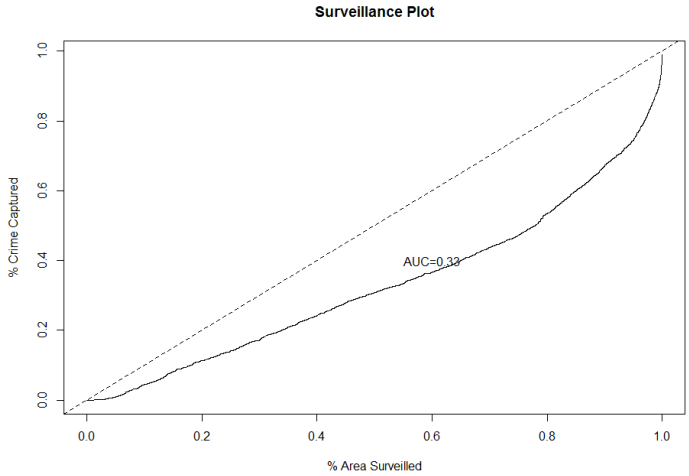


Figure 5: Surveillance plot for RBF kernel SVM

The AUC values are as follows:

5. Logistic regression: 0.678
6. The linear SVM: 0.32
7. The 2-degree polynomial kernel SVM: 0.464
8. RBF kernel SVM: 0.33

---

***Observations***:

1. From the AUC plots we can say that in the following case of using spatial and KDE factors for classification, Logistic regression is performing better than non-linear SVM.
2. As SVM is depends on a small subset of observations and is very robust to the behaviour of observations that are far away from the hyper plane, i.e. only nearby observations have a effect on the class label, in contrast, logistic regression is also less sensitive of the observations that are far away from the decision boundary but it works on the concept of conditional probability and outputs the probability of odds of an observation in either of the classes, which is better suited in the above scenario.

----X----X----