Chicago Crime Rate Prediction

Jin Han | Bryan Li | Wenyu Hu

I Introduction

Criminal activities plague societies by disrupting normal social order, incurring high economic costs on the communities, and causing concerns among residents about their safety. As one of the largest cities in the United States, the city of Chicago also has one of the highest crime rates in the nation.

Our team decided to use socio-economic factors as predictors since they are good indicators of a community's health and vitality. More specifically, the main goal of our project is to predict whether a Chicago community tend to have higher crime rate given the status of its various socio-economic factors.

II Dataset

Datasets we collected include Chicago crime rates[1], socioeconomic factors by communities[2], and population statistics by communities[3]. Since the dataset on socioeconomic factors is for years 2008-2012, we extracted all the crime and population data for the same time frame, and used their 5-year average for the final entry. And crime rate was calculated by counting the number of criminal activities in a given community and scale it according to its total population, so that crime rate is always between 0 and 100.

The final aggregated dataset has 13 independent variables, listed in *Table 1* in the Appendix. We tried both classification and regression, so the dependent variable will be explained in the methodology section. And since Chicago only has 77 communities, the final dataset also only contains 77 instances in total.

III Methodology

Classification

We ordered Chicago communities by their calculated crime rates, and assigned "*Extreme*", "*High*", "*Medium*" and "*Low*" labels accordingly. Since there are only 77 instances, we have unbalanced classes: one class would have 20 instances while the others have 19. After several experiments, we found that different assignment of this "extra" label can actually lead to different models. Thus, we decided to use under-sampling and remove, Edison Park, the community with the lowest crime rate. We also tested that doing this has no effect on the final result.

Since there are 4 classes, the ZeroR baseline has training acc. of 25%. Using Weka, we tried 5 classifiers: Decision Tree, Logistic Regression, Naive Bayes, SMO, and IBK with K=5 . We didn't separate out a test set or even use 10-fold CV because of small sample size. The initial results are shown in *Table 2*, and there is lots of overfitting. Then, to reduce overfitting and high dimension, we performed reduced error pruning on J48, and used meta attributes selection with the others. LOOCV acc. did increase after this, and results are shown in *Table 3*. Also, decision tree with pruning used 3 attributes, *% African American*, *%Asian*, and *Per Capita Income*, and meta attributes selection only kept 2 attributes *% African American* and *% Aged 16+ Unemployed*.

To further improve LOOCV accuracy, we analyzed the Decision tree model (*Figure 1)*, and marked all the incorrectly classified communities on a map (*Figure 2)*. We observed that most of the incorrectly classified communities are from the south side. Thus, we tried adding a new binary attribute that indicates if a community is from the south side or north side. As *Table 4* shows, this significantly increased our LOOCV accuracy. We wanted to use this method to build a good set of attributes, but couldn't find more strong patterns about the incorrectly classified communities. Thus, we decided to try regression on crime rates.

---

[1] Chicago Crime dataset: https://www.kaggle.com/currie32/crimes-in-chicago/kernels.
[2] Chicago socioeconomic factors: https://catalog.data.gov/dataset/census-data-selected-socioeconomic-indicators-in-chicago-2008-2012-36e55
[3] Chicago population statistics, extracted from United States Census Bureau, https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Regression

We first attempted to do regression using all 14 attributes. We tried Multilayer perceptron, linear regression, Bagging+Random Forest and Nearest Neighbor in Weka, and the initial results, given by LOOCV, are listed in *Table 5*. Because of our small sample size, using all 14 attributes also induces the problem of overfitting and high dimension. Also, some of the 14 attributes may be correlated, so that small changes to the inputs can lead to large changes in the model. Thus, we employed some feature selection techniques to limit the number of attributes.

To do so, we ran backward feature selection in R, with *Mallow's Cp* as the criterion. It turned out the best number of features that minimizes *Cp* is 5 (*Figure 3*). These features are: *Hardship*, *%African American*, *%Caucasian*, *ifSouth* and *%Age Under 18 or Over 64*. Using the selected attributes, we got the results shown in *Table 6*. We observed a decrease in MSE with feature reduction, because we got rid of some noise factors.

Among all, Bagging+Random Forest and Nearest Neighbor perform better than the other algorithms. The Bagging meta algorithm reduces overfitting, and so does Random Forest by averaging multiple deep decision trees and training on different parts of the training set. Nearest Neighbor also performs well, possibly because communities with similar socioeconomic factors tend to have similar environment and crime rates.

We reached reasonable correlation coefficient, which means our model can explain most of the variations in the model. However, we achieved somewhat high MSE: our prediction deviate from the actual crime rate by 3.12 on average, which is around 24% of the mean crime rates. This reduces the effectiveness of prediction.

Conclusion: first and foremost, we acknowledge that correlation doesn't mean causation. We can't make causal conclusion on what factors are the underlying cause of high crime rates, but can only say that we observed higher correlation between the selected attributes and crime rates, only in the dataset we aggregated.

IV Future Work

Given the timeframe and scope of this project, we stopped there- however, there are definitely other things we could consider regarding the project:

- First, instead of dividing the city into 77 regions, we could consider dividing the city in other ways. Dividing by Census Tract, for example, can give us more granularity and more examples to work. Data collection, on the other hand, can be challenging.
- Second, we could try divide crime into different categories: violent crimes, property crimes, drug abuses, and etc. Our guess is that they are each correlated with different sets of socio-economic factors, and building separate models for each of them may increase accuracy.

V Appendix

%Households Crowded	%Households Below Poverty	% Aged 16+ Unemployed	%Aged Under 18 or Over 64	%Aged 25+ without Highschool Diploma
%Households with SNAP	Per Capita Income	Hardship Index	% Single Mother Households	
% Caucasian	% African American	% Asian	% Hispanic	

Table 1. Independent Variables

	Decision Tree	Logistic Regression	Naive Bayes	SMO	KNN(K=5)
Training Acc	92.11%	88.16%	61.84%	59.21%	67.10%
LOOCV	67.11%	50.00%	57.89%	48.68%	50.00%

Table 2. Classification: training and LOOCV accuracy. All classifiers are overfitting as their training accuracy is much higher than LOOCV validation accraucy.

	Decision Tree	Logistic Regression	Naive Bayes	SMO	KNN(K=5)
Training Acc	65.79%	60.53%	64.47%	48.68%	64.47%
LOOCV	59.21%	55.26%	59.21%	51.32%	51.32%

Table 3. Classification: training and LOOCV accuracy, after reduced error pruning or meta attribute selection.

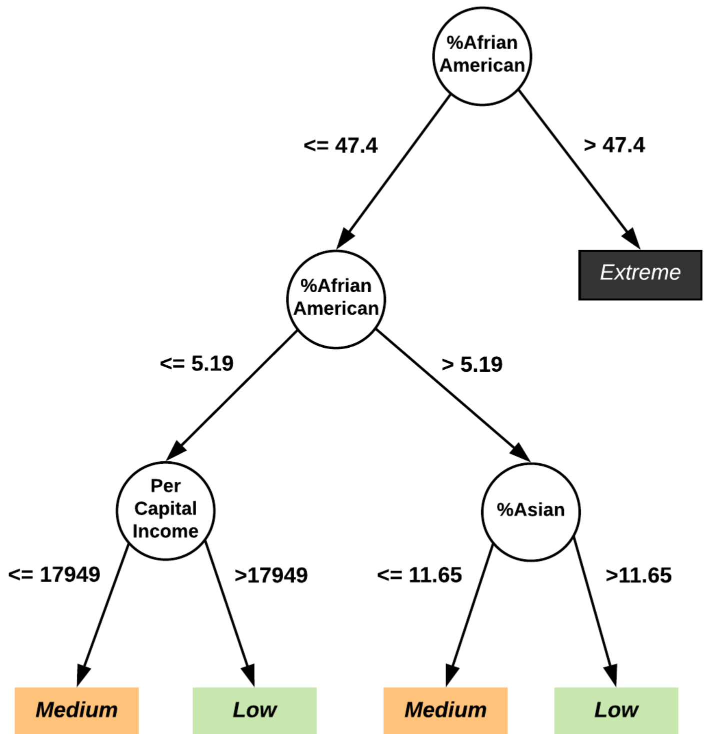

Figure 1. Decision tree with reduced error pruning

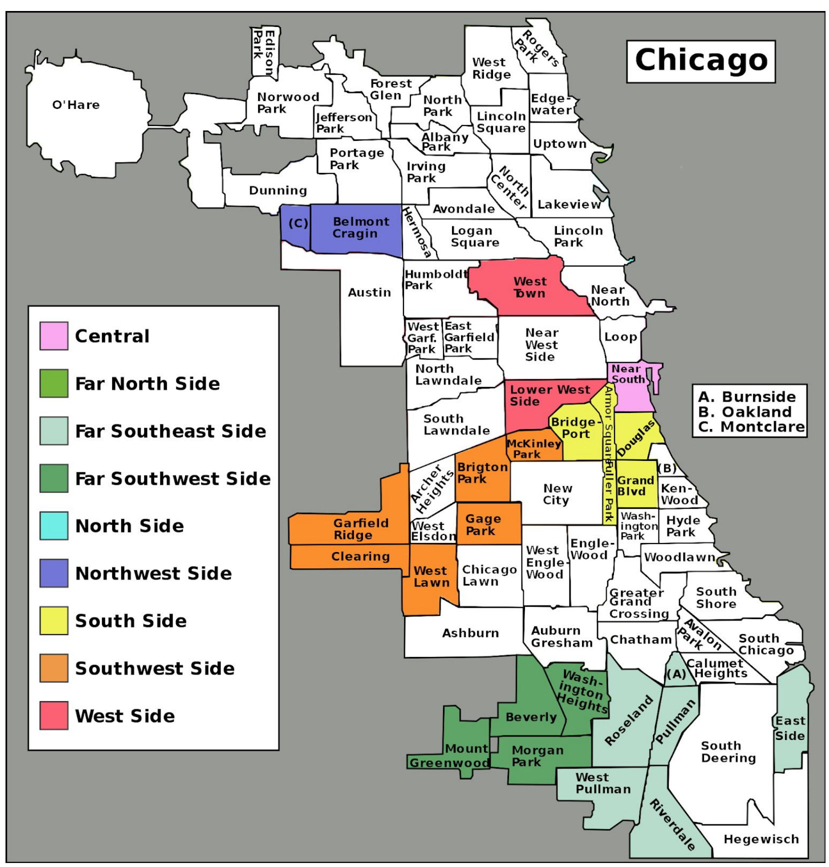

Figure 2. Colored: Incorrectly classified communities. White:correctly classified communities

	Decision Tree	Logistic Regression	Naive Bayes	SMO	KNN(K=5)
Training Acc	73.68%	73.68%	72.36%	68.42%	76.32%
LOOCV	68.42%	69.73%	68.42%	60.52%	68.42%

Table 4. Classification: training and LOOCV accuracy, after adding the new ifSouth binary attribute. Still with reduced error pruning and meta attribute selection

	Multilayer Perceptron	Linear Regression	Bagging + Random Forest	IBk
Correlation coefficient	0.4808	0.6516	0.6848	0.7179
Mean absolute error	4.6998	3.7615	2.9506	2.6984
Root mean squared error	9.0881	5.6109	5.257	5.3079
Relative absolute error	79.8188 %	63.8829 %	50.1108 %	45.8277 %
Root relative squared error	124.5256 %	76.8806 %	72.0315 %	72.7289 %

Table 5. Regression: on Original Dataset without feature selection

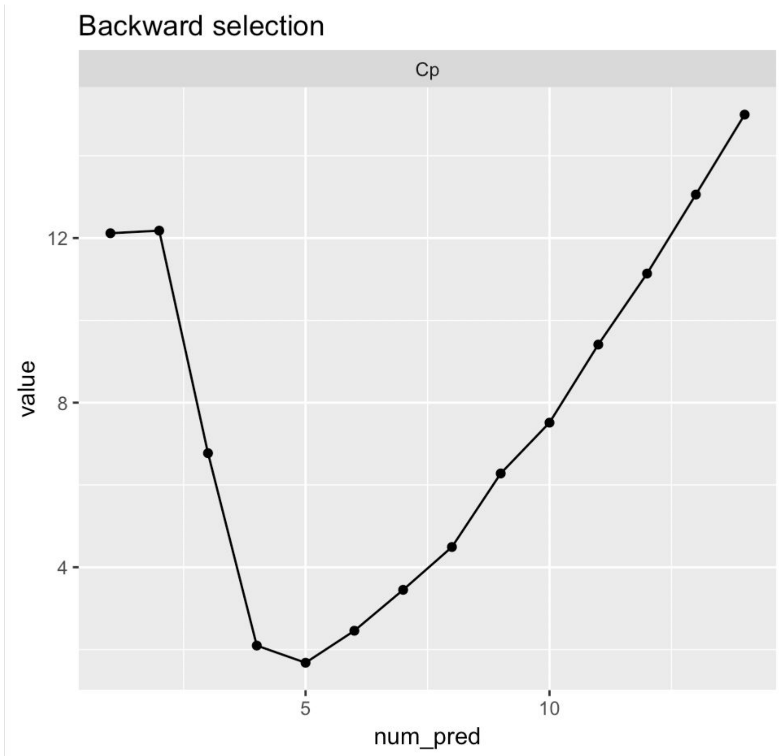

Figure 3. Mallow's Cp with different number of predictors (as can be seen from the graph, #of predictors = 5 gives lowest level of Cp.)

	Multi Layer Perceptron	Linear Regression	Bagging + Random Forest	IBK
Correlation coefficient	0.634	0.7061	0.7117	0.7359
Mean absolute error	3.8731	3.2103	2.7124	2.6649
Root mean squared error	6.0761	5.1398	5.0676	5.132
Relative absolute error	65.7781%	54.522%	46.0664 %	45.2595%
Root relative squared error	83.2541%	70.4262%	69.4358 %	70.3181%

Table 6. Regression: on dataset after feature selection