|  |
| --- |
| Worcester Polytechnic Institute |
| CASE STUDY 4- Crime Data Analysis |
| TEAM 03- Karan, Deepan, Dhaval, Bhakti, Rohitpal |

|  |
| --- |
| DS 501  12-8-2016 |



1. **Introduction:**

A smart city is an [urban development](https://en.wikipedia.org/wiki/Urban_development) vision to integrate multiple [information and communication technology](https://en.wikipedia.org/wiki/Information_and_communication_technology) (ICT) and [Internet of Things](https://en.wikipedia.org/wiki/Internet_of_Things) (IoT) solutions in a [secure](https://en.wikipedia.org/wiki/Information_security) fashion to manage a city's assets – the city's assets include, but are not limited to, local departments' information systems, [schools](https://en.wikipedia.org/wiki/School), [libraries](https://en.wikipedia.org/wiki/Libraries), [transportation systems](https://en.wikipedia.org/wiki/Transportation_system), [hospitals](https://en.wikipedia.org/wiki/Hospital), [power plants](https://en.wikipedia.org/wiki/Power_plant), [water supply networks](https://en.wikipedia.org/wiki/Water_supply_network), [waste management](https://en.wikipedia.org/wiki/Waste_management), [law enforcement](https://en.wikipedia.org/wiki/Law_enforcement), and other [community services](https://en.wikipedia.org/wiki/Community_service).

The goal of building a smart city is to improve [quality of life](https://en.wikipedia.org/wiki/Quality_of_life) by using [urban informatics](https://en.wikipedia.org/wiki/Urban_informatics) and technology to improve the [efficiency](https://en.wikipedia.org/wiki/Efficiency) of services and meet residents' needs.

The business problem to solve:

Providing safety is crucial to improve the quality of life in the cities, and has a significant impact on the promotion of economic growth. Public security is a growing problem for cities worldwide. The world’s cities are bursting at the seams, civic resources are under pressure and crime is harder than ever to police. Safe neighborhood is the one of the fundamental factor for developing city to become a smart city. This is the business problem, to reduce the crime rate in cities.

Why the problem is important to solve?

To promote health, safety and security, it’s important to solve this problem. Reducing the crime rate will have a positive socio-economic impact on the society. It would lead to improve the businesses of some real-estate areas and also help small businesses like grocery stores, restaurants etc. It would have the authorities more aware about certain locations and timings which would help them operate more efficiently.

What is your idea to solve the problem?

Our vision behind this problem statement is to reduce the crime rate in the city to make it safe and smart. In order to reduce the crime rate we need to not only respond to the crime but prevent the crime from occurring in first place and our idea is to predict if the accused is arrested for the given crime type and work effectively to avoid such crimes from occurring where the accused is not likely to be arrested and hence achieve low crime rate in the city.

The goal here is to develop a model based on the crime data that help the authorities to make effective decision and implement effective strategy to keep the city safe and evolve as a smart city.

Crime data analysis is fundamental to effective crime prevention. Knowing as much as you can about crime will help significantly in its prevention.

Determining if particular crimes are increasing; identifying the hot spot locations where crime is concentrated; understanding the temporal trends of offending and analyzing potential reasons for crime trends will be critical features of crime data analysis**.**

What differences you could make with your data science approach?

With the access to large amounts of data and with increasingly smarter statistical analysis, we will be able to use the collected data to foresee and percept various types of criminal acts at a particular time and location, before they even occur.

Such insight will enable city authorities to detect areas of increased crime, which will give them the opportunity to act more deliberately and deploy officers more intelligently, e.g. sending them to the areas that are more exposed to crime. With such systems, police officers will not only respond to criminal acts, but will also be able to act proactively and stop them before they occur.

In order to reduce the crime rate of the city, our solution is not limited to responding to the crime but also to prevent crime from occurring in first place by identifying the crime type that needs more attention by authorities.

We originally categorized the dataset by type, time and location, based on our findings by our problem statements discussed above.

* Crime types: Our model predicts the crime type that requires more attention in order to reduce the crime rate.
* By Locations: Our model predicts the Location that requires more attention for a specific location at specific time, which will help prevent the occurrence of crime and reduce the crime rate.
* By Time: Interesting finding of our model was, the general notion of more crimes taking place during night hours is actually not true. Analysis indicates that most of the crimes occur during day time and during weekdays.

Why do you believe the idea deserves the investment of the "sharks"?

This model can have various positive effects on the life in the city. For everyone,

* For law enforcement authorities-Higher rate of detecting crime, automatic detection of crime, better responsiveness to it, more efficient operations, optimizing budget, etc.
* For cities: Less crime, more attractive for investors, companies and human resources, etc.
* For residents: Safer neighborhoods, faster access to safety information, etc.

We strongly believe, this idea is a step forward in smart city life and will be a win-win situation for everyone. Hence, this idea is a good investment proposal.

What data did you collect? How did you analyze the data?

Our major source of crime data for city of Chicago was from [https://data.cityofchicago.org/.](https://data.cityofchicago.org/)

The website hosts official public record on every crime that is reported to Chicago police department. The dataset does have a limitation with personal records to maintain the privacy. Apart from crime dataset, we also explored in collecting few other datasets required to address and analyses the crime and other related factors that’s could help control the crime to build smart city.

Few other datasets were:

* Police data of city of Chicago
* Real Estate data for city of Chicago
* School dataset

Dataset Size: The dataset includes the crime data from 2001 through 2016, with total observations of 6,193,548.

As shown in the above fig. The crime dataset was categories in to 3 different groups.

* Crime by Time: This helps us analyze when the crime rate is high for a given day or if the crime rate is high during weekdays or weekends.
* Crime by Type: To analyze different types of crimes that occurs in city of Chicago.
* Crime by Location: To analyze the crimes by region and how that relates with our analysis.

For the crime dataset, the total records of 6,193,548 from 2001 – 2016 includes 21 variables.

Upon overview of the dataset, we noticed the data from 2001 to 2016 was significantly skewed over the years.

Data Limitation

There were few limitations with the dataset.

* Data was skewed – the skewed data and was not good for our analysis over years which would lead to bias results.
* Null values: lots of observations had no values or invalid information’s.
* Drop NA
* No Personal info: the data set does not include the victim or accused’s personal info which limited our analysis in going to next level in performing predictive analysis on people and crime.

1. **Define the business problem as a math problem and design a math solution to the problem.**

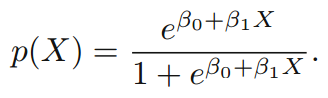
Initially, we thought of applying linear regression to the dataset, in order to predict whether a certain type of crime will observe an arrest of the culprit or not.

(Multiple) Linear Regression:

Y = β0 + β1X1 + β2X2 + β3X3 + β4X4 + …… + βpXp + ɛ

But, we could not use Linear Regression since it regresses on numeric attributes and it requires a numeric target attribute whereas we need a nominal attribute to predict if the arrest occurs or not. Thus, we changed our model to operate on Logistic Regression which can be used to predict nominal attributes.

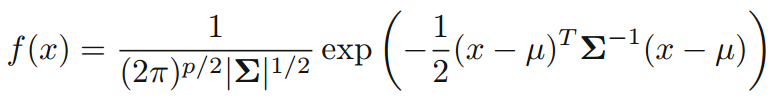
Logistic Regression:



The above equation gives the probability that the given instance belongs to a particular class. Thus, this particular probability helps to predict whether for a particular crime, an arrest has occurred or not. And logistic is best when there are two classes

Linear Discriminant Analysis

Also used this as the target had more than two classes.

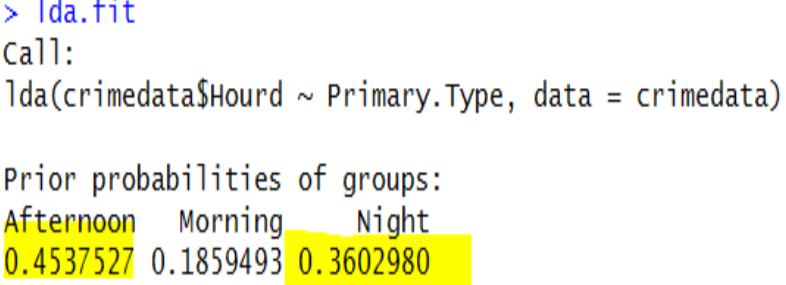


Solution:

The dataset was initially split to training data and testing data based on Year.

Selecting the attributes for logistic regression was another problem tackled. We selected the attributes by applying mixed selection approach. In mixed selection approach, we start with a null set and keep adding variables one at a time. The variables are added until the adjusted R2 value drops below a certain limit. If such a situation occurs, then the least effective variable is removed.

**LDA:**



Above results shows that LDA model correctly classifies the crime type that happen in afternoon and night.

**Logistic:** glm.fit=glm(Arrest∼crimedata$Primary.Type+crimedata$IUCR,data=crimedata, family=binomial, subset=train)

The logistic regression is applied on the attributes ‘Primary Type’ and ‘IUCR’.

Our model was fit and the percentage of correctly classified instances is 86.91%.

We used Confusion Matrix, Sensitivity, Specificity and Precision as performance parameter.

### **The Hacking Part**

Data Collection

As explained earlier we collected data from city of Chicago website. This data consists of few columns to play on. So, we decided to add additional columns by transforming Timestamp column into Date, Day, Month, Year, Time of the day (Morning, Afternoon, Night). Also, we extracted Zip Codes using geolocator library in python using latitude and longitude given in data set.

Moreover, we decided to work on Restaurant data set. We extracted rating from the yelp, classified based on latitude and longitude, we merged the restaurant ratings based on the location. This information was merged with crime dataset to suggest better and safe restaurant at a given time.

We have attached the screenshot on the interactive visualization which helps user to get insight about crime data set and using which he can make decisions to reduce the crime. The visualization are performed on the columns which are highly correlated to our targets.

Implement a small Demo/Prototype/experiment result figures for the "product" of your data science company:

### Dashboard

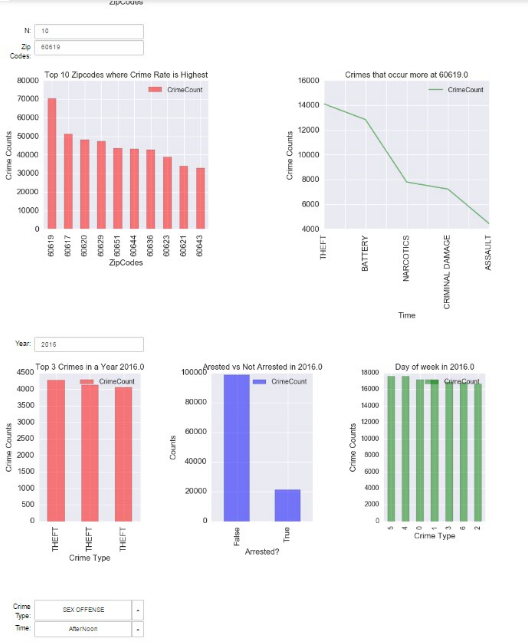
We also developed interactive UI, where the user can narrow down the by entering more details to narrow down the analysis.

For instance,

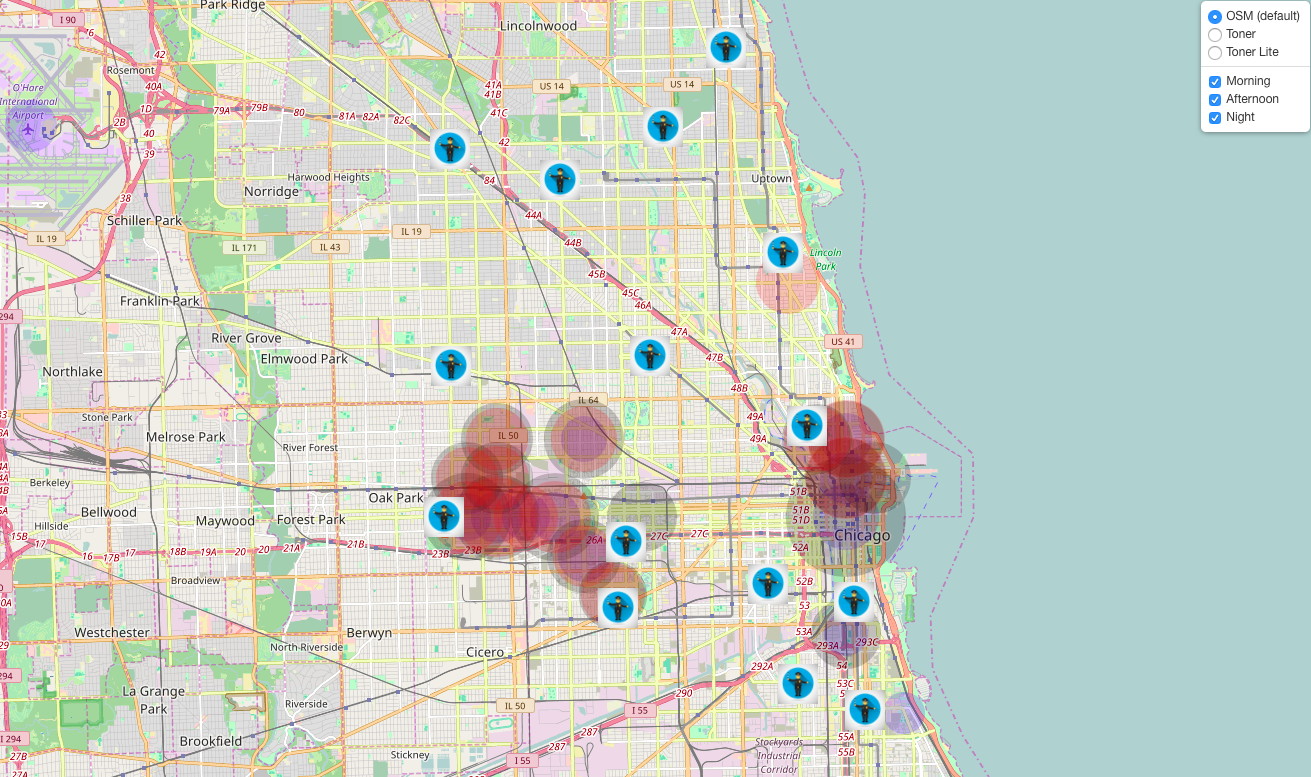
* Get top 10 zip codes with highest crime rate.
* Enter desired zip code – to get top crime types
* Enter desired crime type – to get the occurrence by time

This interactivity clearly guides you to what type of crime occurs at given location and gives you a summary on when the crime is likely to occur, these interactive UI would help authorities to narrow down the analysis for specific location, time (Day, Week, Month).





An attempt to plot the analysis on the map, to visually analyze and findings on crime data against the CPD dataset. This map helps us to find the relationship between crime location and police stations in specific areas.

****