CRIMES IN ATLANTA, GA

Spring 2018

Crime reports collected by Atlanta PD in 2017

In this project we evaluate the crime data collected from Atlanta Police Department and examine the effects of several location and time-based parameters on the occurrence and type of crime.

Crimes in Atlanta, GA

CRIME REPORTS COLLECTED BY ATLANTA PD IN 2017

CONTENTS

MOTIVATION

Atlanta, the capital and most populous city of the state of Georgia, is the 6th most dangerous city in the country according to this report. The rate of crime is more than twice of the national average, with 5203 violent crimes and 25,556 property crimes reported in the year 2015. Our motivation for deciding to work on the data of crimes committed in Atlanta in the previous year was to analyze the distribution of crimes by region, time, and type of crimes committed, among other variables, and try to develop a predictive model towards predicting the type of crime that might be committed given the variables stated above.

THE DATA

The data has been collected from Atlanta Police Department's website. It consists of 26,760 data points, i.e., instances of recorded crimes, and 23 columns consisting of fields related to occurrence of crime, type of crime, location codes, neighborhood in which the crime took place, etc.

CLEANING THE DATA AND FEATURE SELECTION

The data consists of several NaN values and "Unk" values in fields. We begin by removing these values to obtain a consistent dataset. Further, variables like Location Type, Beat no., etc. will be removed for ease of model building. Other variables like Apartment number will be removed because they contain a lot of missing values.

The neighborhood variable consists of 236 different classes initially. For this analysis we will be focusing only on the top 25 most crime-ridden neighborhoods, as we wish to prioritize the densest regions first, as well as for ease of computation.

EXPLORING THE DATA

Using various plots, we analyze the spread of the data according to different predictors. We find that Downtown is the densest area when it comes to crimes, followed by Midtown and Old Fourth Ward. When split by type of crimes, Larceny From Vehicle is the most frequent crime, with close to 4500 instances in the cleaned data with the Top 25 neighborhoods alone.

With respect to days of the week, Saturday ranks first, while looking at the month-wise distribution we see that most number of crimes take place in October. The frequency of crimes increases towards the end of the day, and peaks at 7 p.m.

This data is extremely important, as it helps us understand which variables affect the rate of crime and type of crime occurring in Atlanta. We will be making use of these inferences for our predictive modelling.

IMPLEMENTATION

To begin our predictive modelling, we consider the different classification models possible:

1. Multinomial Logistic Regression
2. Linear Discriminant Analysis
3. Quadratic Discriminant Analysis
4. K-Nearest Neighbors
5. Random Forest
6. MART
7. Artificial Neural Networks

Multinomial Logistic Regression:

This is logistic regression extended for response with more than two classes.

The log-odds for this method is defined as:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Using a bit of algebra, we can calculate the probability that a data point will belong to a particular class as follows:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Linear Discriminant Analysis:

LDA approximates the Bayes' Classifier – the Gold Standard for Classification – by assuming that the observations are normally distributed. Thus, the probability density function for observations will be:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

LDA trains the model to create a decision boundary, which can then be used for testing. The Linear Discriminant function is:

$$\hat{\delta}_k(x) = x\cdot\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

We also assume that the classes have equal variance, i.e., homoskedasticity. LDA is more popular for response variable with more than two classes, so we expect a better accuracy from the model if our assumptions hold true.

Quadratic Discriminant Analysis:

QDA works in a similar way as LDA, the only difference being that the Discriminant function is quadratic in X:

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \\ &= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\pi_k\end{aligned}$$

QDA assumes that the classes have a different variance of their own, i.e., heteroskedasticity. QDA will perform better than the linear classifiers discussed above if the decision boundary is non-linear.

However, for this data, we find that the number of data points are too low to train a QDA model. Hence, we move on to other models.

Random Forest:

Random Forest is a tree-based method which uses bootstrap aggregation while simultaneously changing the predictors considered for each bootstrap sample. For e.g.: we consider a random 'm' predictors out of the total 'p' (generally, m = sqrt(p)) for each sample. This helps us decorrelate the samples and reduce variance. We build 100 trees for bootstrapping. Random Forest performs well for classification with multiple response classes and different number of observations per class.

Multiple Additive Regression Trees (MART/ Gradient Boosting):

The accuracy of a predictive model can be boosted in two ways: Either by embracing feature engineering or by applying boosting algorithms.

MART is an implementation of the gradient tree boosting methods for predictive data mining (regression and classification) described in Greedy Function Approximation: A Gradient Boosting Machine and Stochastic Gradient Boosting. Gradient boosting is an ensemble supervised machine learning model that builds up the concept of the random forest algorithm. The algorithm iteratively builds trees that minimize the error, and thereby descends towards an optimal set of predictive trees. Already learned trees are kept, and new trees are added one after another to minimize the objective function (error in predictions). The trees are grown sequentially: each tree is grown using information from previously grown trees. Each tree is fit on a modified version of the original data set based on the previous trees built.

The trees are accompanied by a regularization parameter to avoid overfit.

Artificial Neural Networks:

As defined by Wikipedia, An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The novelty of neural network lies in the way it learns from the given input trends to predict output.

A neural network typically consists of 3 layers: input layer, output layer, and hidden layer which converts the input in a form which is readable or useable by the output layer. The

node multiplies each of the inputs by some weight and the sum is then passed to an activation function. The following equation gives the output function:

$$f(\mathbf{x},\mathbf{w}) = \phi(\mathbf{x}\cdot\mathbf{w}) = \phi\left(\sum_{i=1}^{p}(x_i \cdot w_i)\right)$$

x: input vector, w: weight vector of the neuron
p: number of inputs to the neuron, ø: activation function

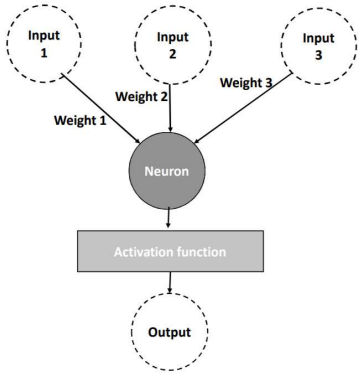


Figure 1. Artificial Neuron

We have used sigmoid function. Data is normalized before analyzing. Only the important predictors are considered for analysis and the response is coded as binary variable with 11 levels.

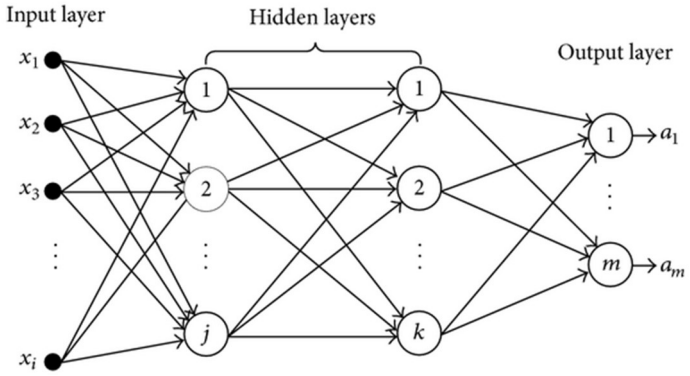


Figure 2. ANN Architecture

RESULTS

METHOD	PREDICTION ACCURACY
Multinomial Logistic Regression	41.33%
LDA	41.43%
Random Forest	41.35%
MART	44.15%
Artificial Neural Networks	49.91%

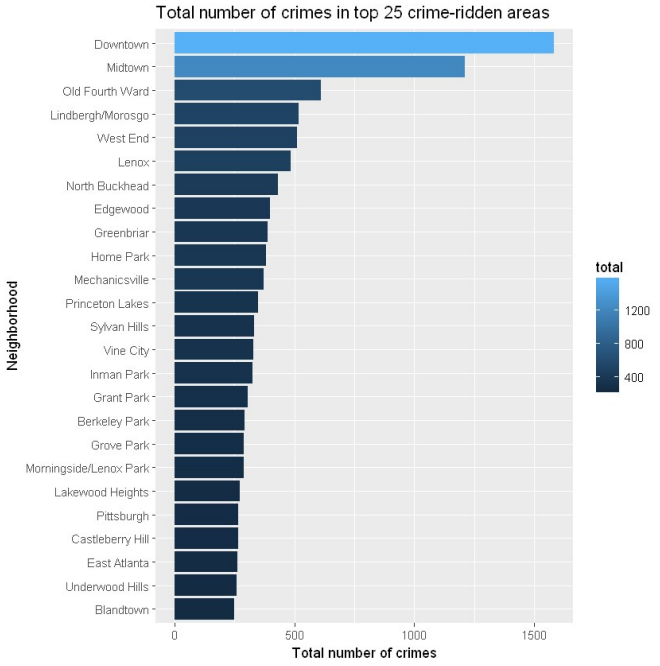
CONCLUSION

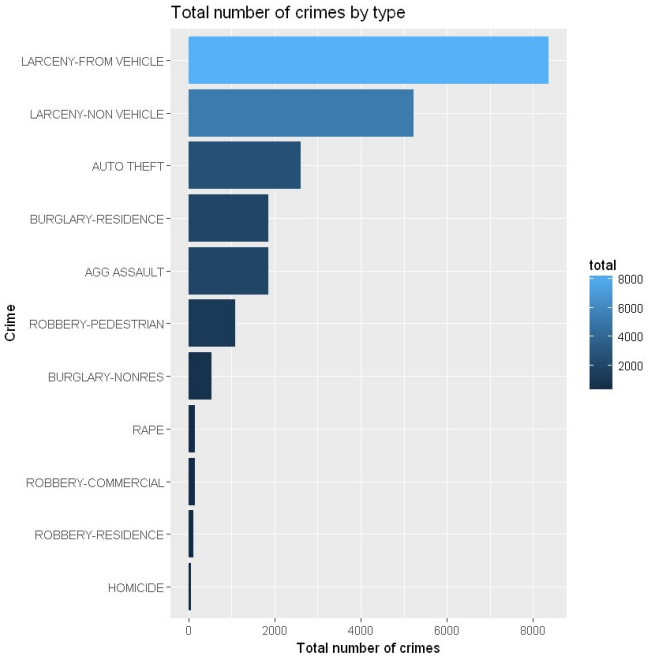
The data shows that thefts from vehicles are the most frequent types of crimes in Atlanta, Georgia. Using this and the other data exploration plots, we can help Atlanta PD reinforce their preventive measures.

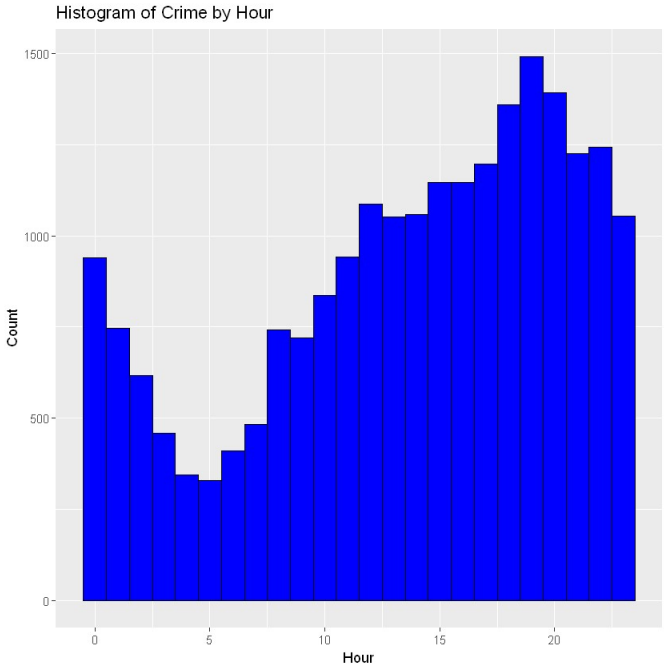
The analysis and our models can be used to predict what type of crime will be committed based on the time and location input to the model. This information can be of great help when it comes to public awareness and preventive action.

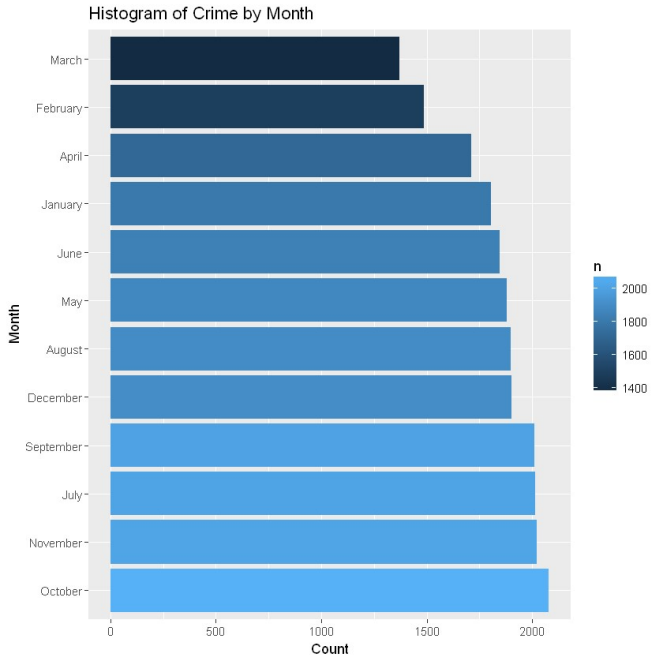
Our models can be made more accurate by implementing a time-series model, which would help us analyze the trends and seasonality factors more accurately.

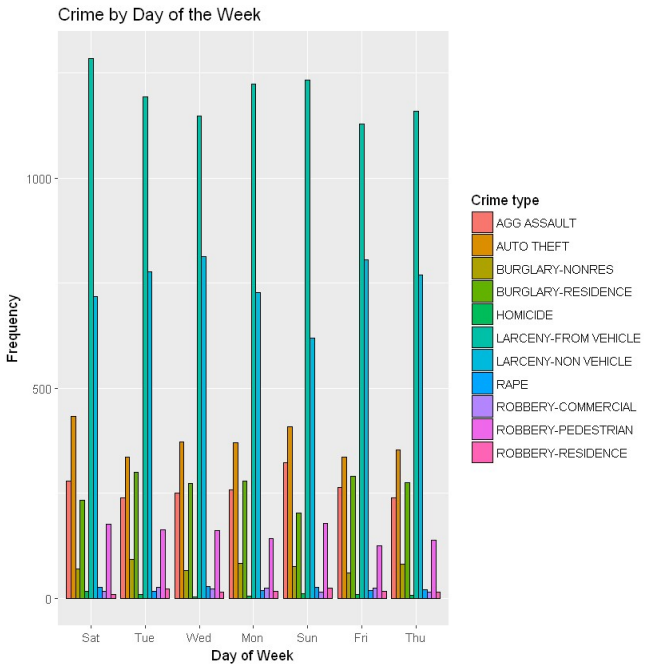
Appendix: Plots generated for data exploration

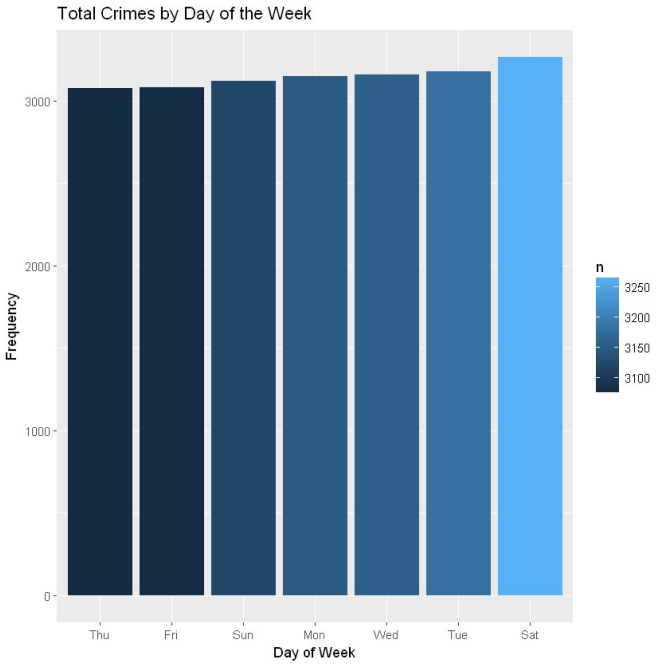









Crime by Day of the Week
Frequency
1000
500
0
Sat
Tue
Wed
Mon
Sun
Fri
Thu
Day of Week
Crime type
AGG ASSAULT
AUTO THEFT
BURGLARY-NONRES
BURGLARY-RESIDENCE
HOMICIDE
LARCENY-FROM VEHICLE
LARCENY-NON VEHICLE
RAPE
ROBBERY-COMMERCIAL
ROBBERY-PEDESTRIAN
ROBBERY-RESIDENCE


Total Crimes by Day of the Week
Frequency
3000
2000
1000
0
Thu
Fri
Sun
Mon
Wed
Tue
Sat
Day of Week
n
3250
3200
3150
3100