# Preventing Urban Crime

*Jonathan Posada, Calvin Ying*

*ORIE 4741 Final Report*

## 1 Abstract

Numerous reports have shown that trees have a direct relationship with happiness. Take Costa Rica as an example. Having more than half of its territory covered by rainforest, Costa Rica has consistently been celebrated for being one of the happiest and greenest countries in the world. The 2016 Happy Planet Index compiled by the New Economics Foundation (NEF) placed Costa Rica as the happiest country in the world, judging from factors such as life satisfaction and ecological footprint. [1]

Observing that trees can have a direct impact on happiness, our group set out to explore the correlation between trees and crime. Using crime and tree data from OpenDataPhilly [2] [3], we strive to predict whether planting more trees would decrease the number of crime incidents in Philadelphia, PA. The motivation for the study is that increased tree planting could provide a simple and inexpensive method for reducing crime. This is in comparison to most accepted solutions to crime such as increasing alcohol taxes, providing affordable housing or changing police tactics [5].

For this project, we built two regression models, one that includes tree count as a predictor of crime and another that does not consider tree count. We compared the out of sample error for both models: if the model with tree as a predictor has lower out of sample error, we would conclude that the model with tree can better predict the crime level in Philadelphia. We

expect our findings to provide the Philadelphia Police Department and Parks and Recreation Department a tool to decide whether to plant more trees in specific neighborhoods, allowing them to take proactive steps to reduce crimes throughout the city.

## 2 Exploratory Data Analysis

The original crime data set is collected from 2006 to 2016, it has one million rows (observations) and 14 columns (features). The data set is as follows:

| Dc_Dist | Psa | Dispatch_Date_Time | Dispatch_Date | Dispatch_Time | Hour | Dc_Key | Location_Block | UCR_General | Text_General_Code | Police_Districts | Month | Lon | Lat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 25 | 4 | 6/12/2015 19:54 | 6/12/2015 | 19:54:00 | 19 | 2.02E+11 | 3300 BLOCK N 05TH ST | 1800 | Narcotic / Drug | 18 | 2015-06 | -75.13797 | 40.002517 |
| 35 | 2 | 4/8/2015 4:02 | 4/8/2015 | 4:02:00 | 4 | 2.02E+11 | 5200 BLOCK WESTFORD RD | 2100 | Driving Under | 20 | 2015-04 | -75.116068 | 40.028974 |
| 24 | 2 | 10/6/2015 1:35 | 10/6/2015 | 1:35:00 | 1 | 2.02E+11 | 100 BLOCK E LIPPINCOTT ST | 300 | Robbery | 17 | 2015-10 | -75.129454 | 39.997584 |

Table 1

As seen in the image above, some of these features are redundant as date_time, date, time and hour are all features that encode the time of the incident. Therefore, we decided to only look at the "Hour" column when we conducted our analysis. The Text_General_Code column shows the different crime types, but some of these crime types can be grouped together to simplify our analysis. For example, "Burglary Residential" and "Burglary Non-Residential" can be grouped under the same category, "Burglary." We also omitted crime types with general names, such as "All Other Offenses" and "Other Assaults."

In addition, we decided to append a new column "Zip Code" for each incident of crime. This allows us to generalize the data and provide useful insights across different neighborhoods. We used the Google Maps Geocoding API [4] to convert the longitude and latitude coordinates of a crime incident to the corresponding zip code. As a final step, we deleted any rows with missing values. This meant removing 0.8% of samples which would not hugely impact our analysis.

After cleaning the data, we plotted a time series graph (Figure 1) to show the trend of crime incidents. We later decided to only use data from 2015 because the data is the most recent and complete. We designated 80% of the data as the training set and 20% of the data

as test set.

After we have cleaned the data, we performed feature transformation to create a new data set with each zip code having its own row. The new data set is shown below in Table 2:

| Zip Code | Paid Employee | Median Household Income (K) | Median Age | Population Density (per squared miles) | Crime Count | Hour_0 | Hour_1 | Hour_2 | Hour_3 | Hour_4 | Hour_5 | Hour_6 | Hour_7 | Hour_8 | Hour |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19102 | 29670 | 59 | 32.3 | 28302 | 60 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | |
| 19103 | 77089 | 49 | 34.5 | 33364 | 85 | 3 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | |
| 19104 | 75000 | 19 | 23.7 | 16869 | 201 | 7 | 8 | 12 | 7 | 1 | 4 | 5 | 5 | 6 | |
| 19106 | 22624 | 78 | 37.5 | 17131 | 49 | 2 | 1 | 2 | 2 | 0 | 0 | 1 | 1 | 3 | |
| 19107 | 40445 | 37 | 31.1 | 28187 | 109 | 3 | 5 | 2 | 3 | 0 | 0 | 0 | 2 | 10 | |
| 19111 | 10402 | 44 | 36.5 | 11468 | 133 | 6 | 3 | 6 | 6 | 2 | 1 | 2 | 3 | 2 | |

Table 2

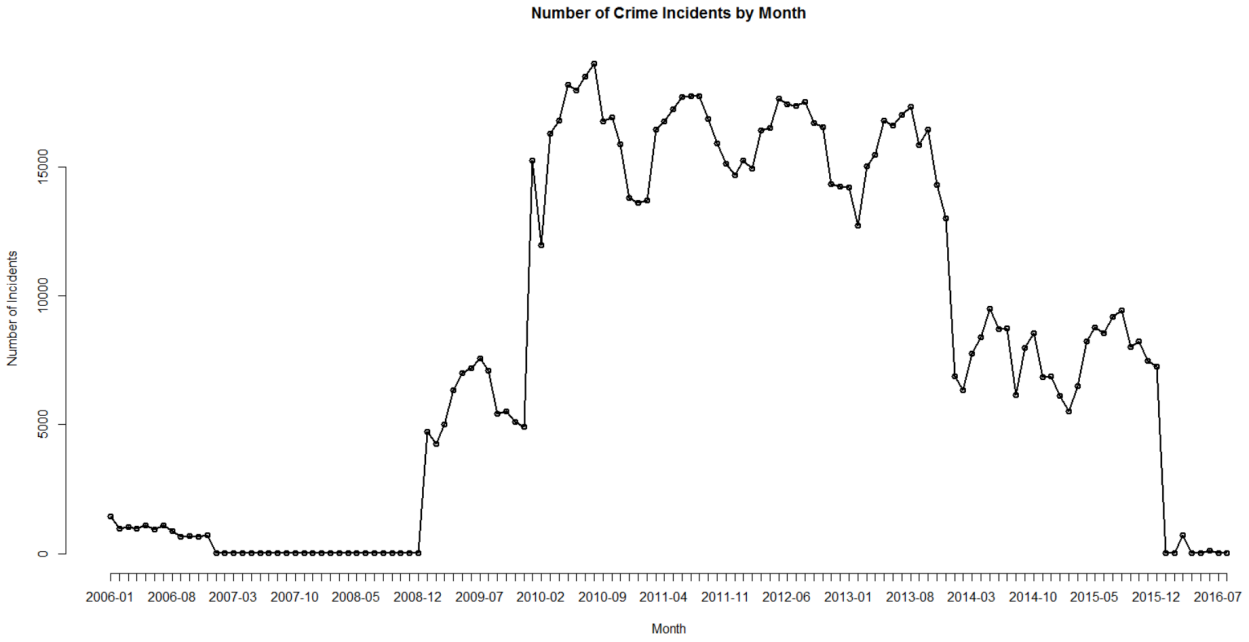


Figure 1: Number of Crime Incidents in Philadelphia by Month from 2006 to 2016

As seen in Table 2, we transformed the "Hour" column in the original data set to 24 columns in the new data set, each column representing the total number of crimes that occurred in that hour. We also summed up the number of crimes that happened in each zip code using the original data set and created a "Crime Count" column in the new data set. Several new features were added to the new data set as well. Hoping to find some predictors of crime we added number of paid employees, median household income, median age and

population density for each zip code. At last, we appended a "Trees" column to our data set. The new data set has 47 rows and 31 columns, which we used to create regression models.

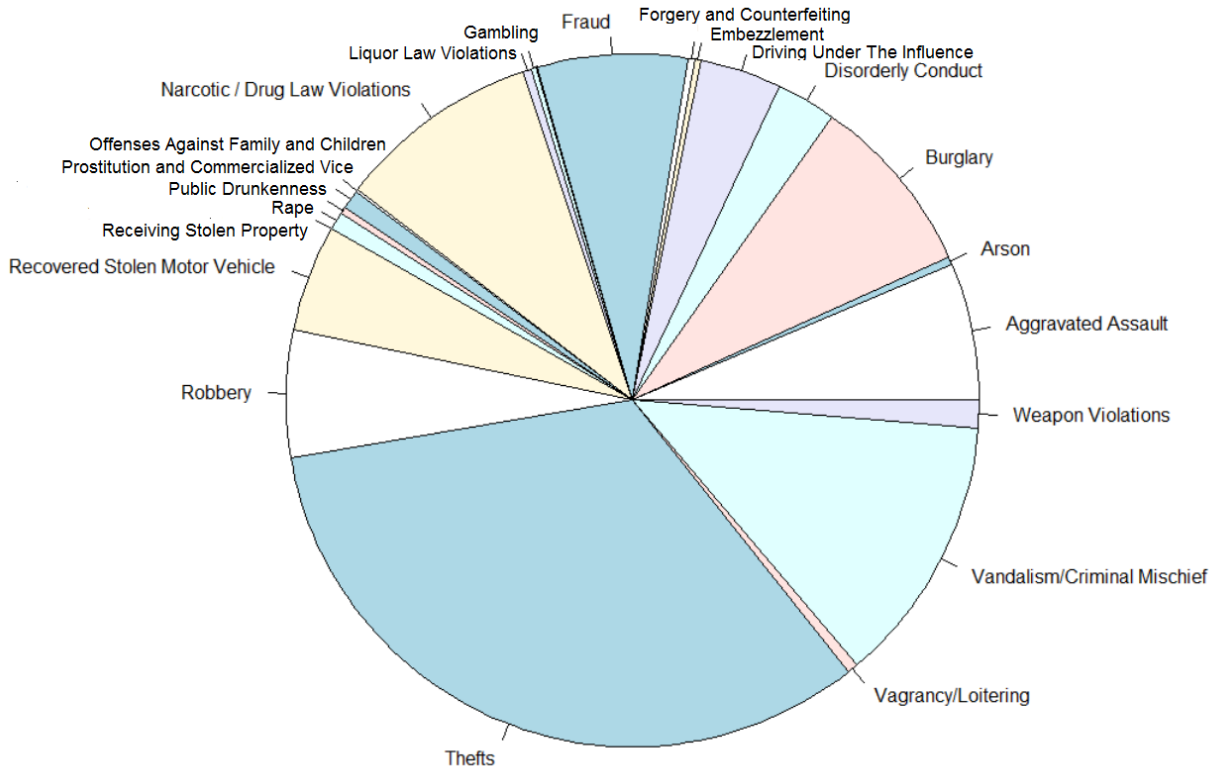


Figure 2: Crime Types

# 3 Learning a Model

The first model we applied was a simple linear regression on the training set without tree data. Using the aggregate number of crimes as the output vector y, and four feature vectors as X, we created the following model.

This model is not surprising and indeed confirms the authors' prior intuition. Higher employment rates, higher income, and an older population all contribute to reducing crime. However, it is interesting to note that a more densely populated city contributes to more crime.

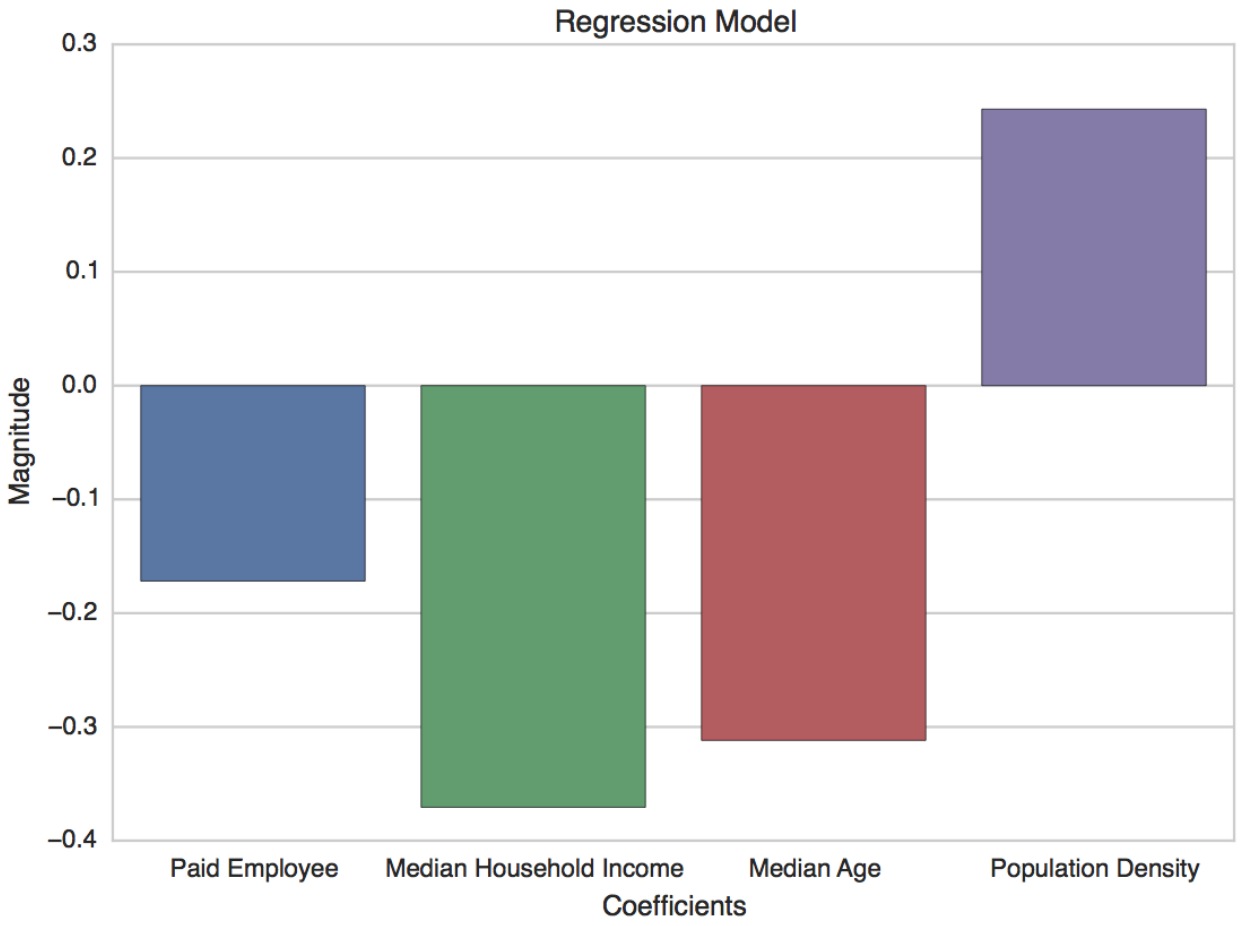


Figure 3: Regression Model

With an initial model working, we generated additional models and then evaluated each model to select the winning model to be used in practice. Because our data contained 47 rows and 31 columns, our simple linear regression model performed well but it was probably overfitting the data. To solve the overfitting problem, we used regularization techniques. We expected many factors to contribute to crime (and thus we do not want sparsity in the data) – indeed more than just the four features examined here – so we used Ridge Regression as the second model. We used cross-validation to determine the alpha parameter of 10. For the third model we used Lasso to evaluate our previous assumption that many factors contribute to crime; if the authors are correct then a Lasso model would perform poorly. Below is a figure comparing the 3 models by the in-sample and out-of-sample error.

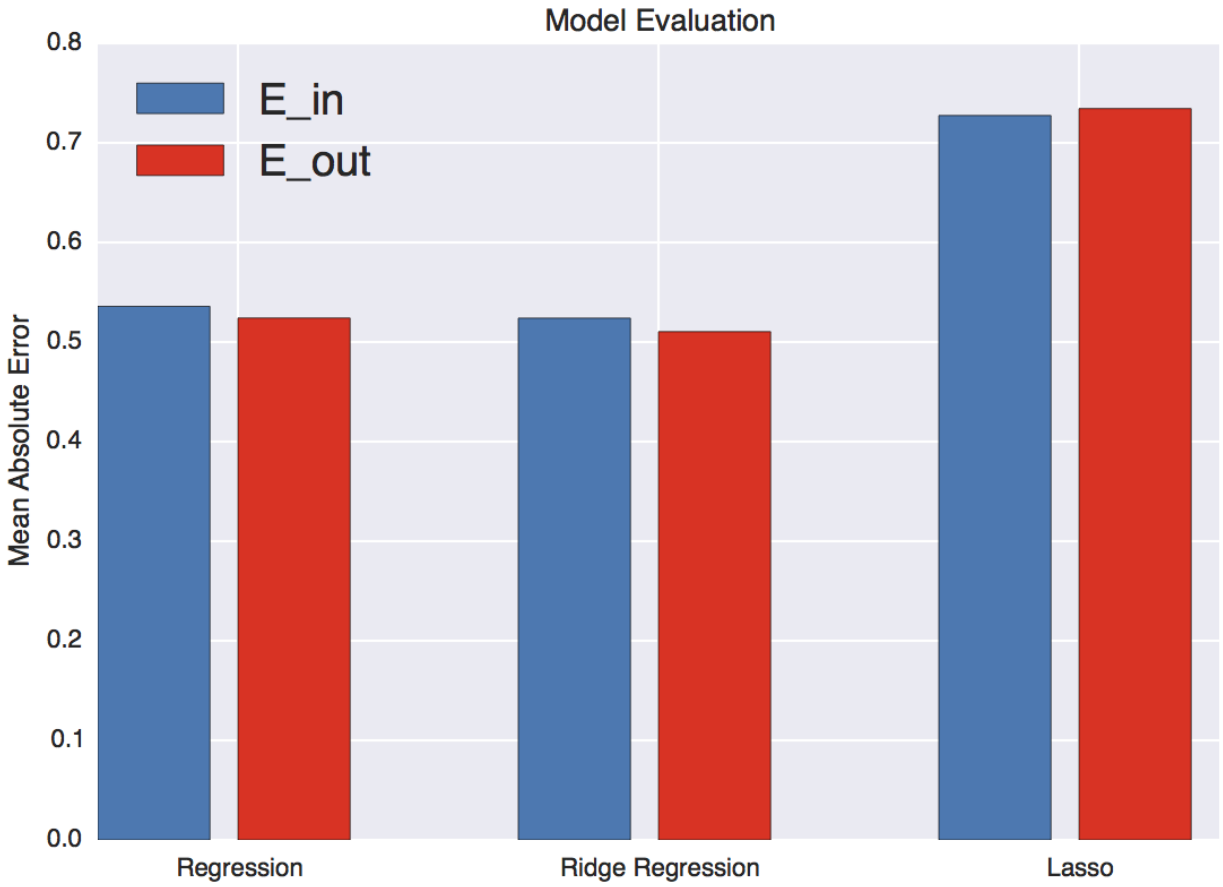


Figure 4: Comparing Models

To compare the three models we used the mean absolute error as the scoring metric. Since aggregate crimes is a large integer value, using an absolute difference was the obvious choice as a metric. As we see in figure X, Lasso performed worse than the other two models, confirming the assumption that crime has numerous predictors. We also note that Ridge Regression performed the best showing that regularization improved upon the original model.

Now, we append the data set with the trees data and determine how trees affect crime. We use the same three learners for generating models and again Ridge Regression performs the best, now with alpha=20. Below is a comparison of the three models and the winning model.

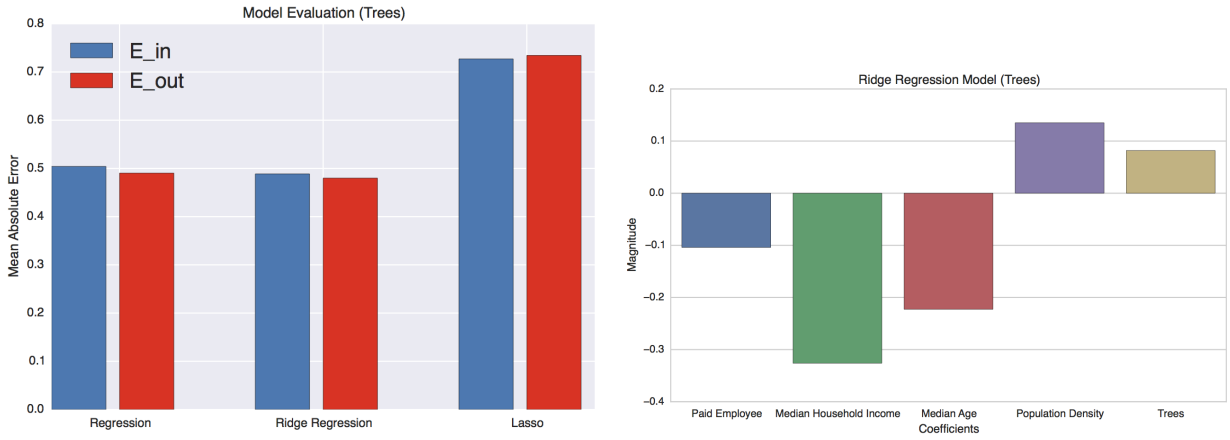


Figure 5: Comparing Models

Surprisingly, more trees in a community contributes to higher crime. This seem unintuitive at first but the addition of trees also introduces more shaded area which becomes dark at night. Studies have shown that improving street has a significant effect in reducing crime [6].

# 4 Conclusion

Ultimately the study provided a couple interesting insights. We confirmed some prior assumptions about the relationship between jobs and age and the crime rate. However, we were surprised to discover that more trees in Philadelphia contributes to higher crime. However, we used total tree count as a feature in the analysis and using other features such as tree density may have produced different results.

The biggest challenge we encountered in this project is finding relevant data sets that we could use. Since our original data set lacked interesting features, we had to find additional data sets to append to our original one. However, it took us a long time to join different data sets together. In future work, if time permits, we would like to include more data and add additional feature vectors. We were forced to work with fractions of the total available data since most samples were logged using geographic coordinates. These samples needed to be converted to zip codes using the Geocoding API which has a maximum of 2500 free

requests per day, making it difficult to work with big data. We would also like to include trees data over time and see how the fluctuations in tree count affects the crime rate. More feature vectors could also improve our model and determine more predictors of crime.

# References


[1] Happy Planet Index

http://happyplanetindex.org/countries/costa-rica

[2] Philadelphia Crime Data

https://www.opendataphilly.org/dataset/crime-incidents/resource/d6369e07-da6d-401b-bf

[3] Philadelphia Street Tree Inventory

https://www.opendataphilly.org/dataset/philadelphia-street-tree-inventory

[4] Google Maps Geocoding API

https://developers.google.com/maps/documentation/geocoding/intro

[5] Ways to Reduce Crime

http://www.urban.org/urban-wire/five-ways-reduce-crime

[6] Street Lighting and Crime

https://www.crimesolutions.gov/ProgramDetails.aspx?ID=84