CS 229 - Final Project Milestone

Kane, Alyson
alykane@stanford.edu
SUID #06079278

Sagalovsky, Ariel
asagalov@stanford.edu
SUID #05519028

November 20, 2016

1 Introduction

The purpose of our project is to gain insight into crime patterns across major U.S. cities. We are interested in understanding which demographic factors may contribute to neighborhood-level crime and if these factors are consistent across various cities. We will implement both supervised and unsupervised learning techniques to build predictive models and identify key attributes that contribute to crime rates. Project progress and up-to-date code can be found at the following path: https://github.com/asagalovsky/CS229-Project

2 Data Sources

Two main data sources were compiled for our project: (1) 2010 incident-level crime data, and (2) 2010 American Community Survey (ACS) census data. Both data sources were collected for 6 cities across the US: New York City, Chicago, San Francisco, Detroit, Philadelphia, and Washington, D.C.

Using the OpenData portal hosted by each city's local government, we downloaded all crimes occurring in 2010. Each dataset includes date and time of occurrence, type of crime, and geographic identifiers (latitude/longitude). Census data was downloaded at the tract-level from the 2010 100% ACS survey. For this milestone, we have included demographic data (age, gender, race) and housing data.

3 Data Processing

The majority of the effort to date has been spent collecting and aggregating data. Because the locations of crime reports are generally reported with latitude/longitude coordinate identifiers, we had to find a way to map the locations to their respective Census tracts.

In order to accomplish this task, we downloaded all the necessary shape files for each county in our study. These files contain exact latitude/longitude coordinates of the polygons defining each Census tract in the county. We used the `sp` package in R to overlay the borders of the Census tracts on top of the map of the locations of each crime incident. In doing so, we were able to tag each of observations in our crimes dataset with the appropriate Census tract. Finally, we aggregated the counts of crime incidents at the tract level, as we ultimately hope to predict crime rates at this level.

The Census dataset also required a fair bit of pre-processing. We scrubbed the raw dataset to retain only relevant fields. For instance, individual tracts were tagged with an 11-digit numeric ID, which we converted to the appropriate value, generally between one to four digits, sometimes with two decimal places. We used a very systematic approach in verifying that each city had all tracts correctly tagged and in the same format, to allow for a more straightforward merge across datasets later on. Further, we used our domain expertise to reduce the number of features from around 300 to 50. To allow for better comparisons across tracts, fields

with raw numbers were dropped and only percentages were kept.

Once the individual city-wide crime incident datasets were cleaned and aggregated, we stacked them into a single data frame before merging with the Census data. Crime rate itself, which will be our response variable, was a field we created by dividing the total crimes for each tract by its population.

4 Initial Results

Our methodology to date has been relatively straightforward and has been mainly for proof-of-concept and sanity checking. We began by normalizing each predictor take on mean zero and variance one, while the response (crime rate) was transformed using logarithms for stability of predictions. Figures 1 and 2 below depict the response variable prior to, and after, transformations.

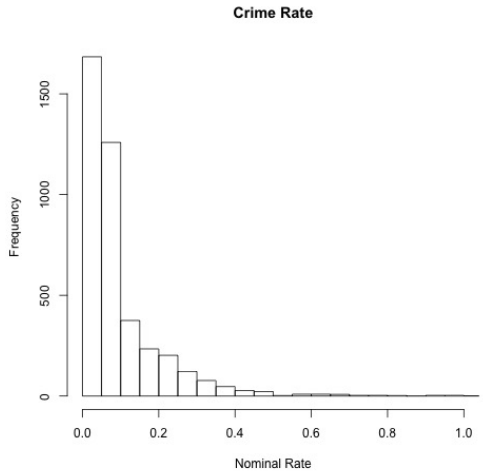

Figure 1: Original Response

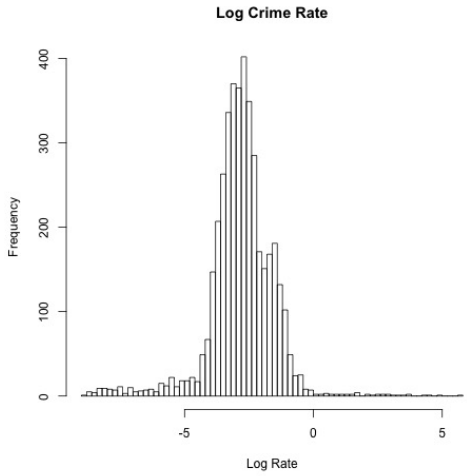

Figure 2: Transformed Response

Before implementing our most basic supervised learning models, we split the merged dataset into two randomly chosen subsets for training (70%) and testing (30%). We built each of our models on the same training set and measured our predictive error on the held-out test set.

At this stage, we have mostly explored linear models (OLS, Ridge, Lasso, and Elastic Net). After tuning the `alpha` parameter in the Elastic Net model, we can report the following errors (RMSE) on the held-out test set:

Model Type	RMSE
OLS	0.949
Ridge	0.898
Lasso	0.906
Elastic Net	0.903

We seem to have comparable results for each of the penalized regression models at the moment and the resulting predictions are similarly distributed. Below are some snapshots of one of our model's residuals:

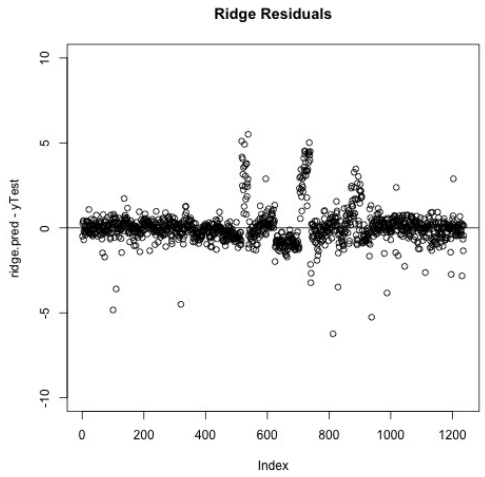

Figure 3: Ridge Residuals Scatterplot

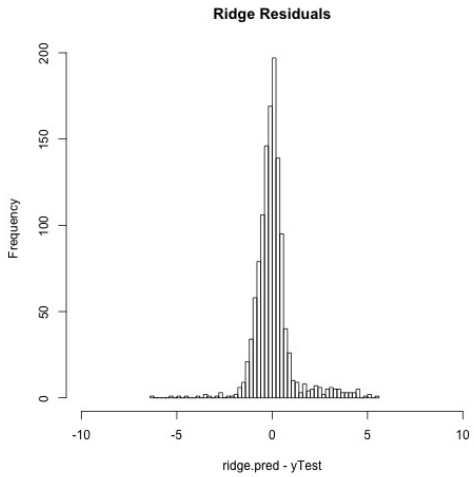

Figure 4: Ridge Residuals Histogram

We suspect that the relationship between the response and the predictors in our data is non-linear. In a first attempt to implement decision trees, we ran a Random Forest model with 500 trees and 10 variables considered at each split point. The resulting error was considerably lower than the linear models previously considered: 0.768. This is the biggest reduction in RMSE error we've seen thus far, confirming our hypothesis that non-linear methods will likely outperform linear methods. We will continue to tune parameters in decision trees and explore other non-linear methods, further detailed in next steps. Random Forests are also useful in analyses with a large number of predictors, because this method helps choose which variables are most important at each node. See below for a variable importance plot:

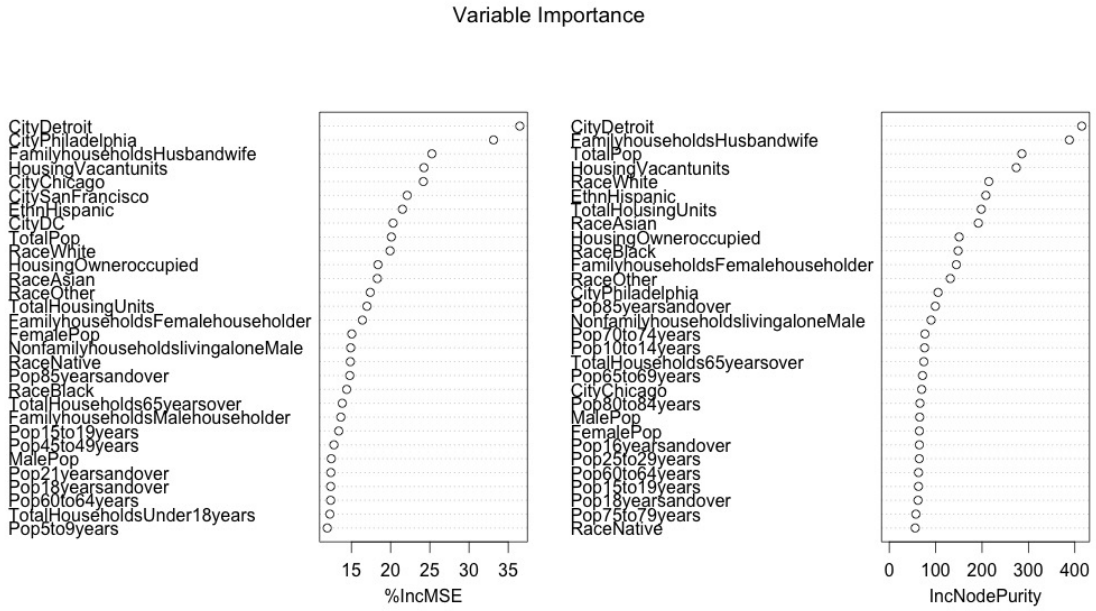

Figure 5: Random Forest Variable Importance

5 Next Steps

After reviewing initial results of our analysis, there are a few more avenues that we would like to explore. Currently, all crimes are being considered equally. That is, we are aggregating all types of violent and non-violent, white- and blue-collar, and misdemeanor and felony offenses into a single group. In subsequent iterations, we wish to explore separate models for violent and non-violent crimes. We have reason to believe that demographic data, specifically, will lead to an increase in predictive accuracy after splitting the crimes by type.

Also, we may wish to vary our feature set to include more demographic and geographic factors. For one, looking at socio-economic indicators such as median housing price, education attainment levels, and poverty levels may be far better predictors than the features we are currently using. Additionally, there have been studies in the past that have strong statistical evidence to support the hypothesis that cold weather leads to a decrease in crime rates. It may be of use to bind regional weather statistics to our dataset for each city in an attempt to understand whether this claim is well-founded.

Further, we have already seen that a non-linear modeling approach may be best for this prediction problem. We wish to expand our efforts in fine tuning parameters for our Random Forest model, as well as attempting other tree-based models such as Gradient Boosted Machines. We imagine that the best model for prediction will be an ensemble of several methods, not one single model.

Finally, part of our goal with this project is to understand which factors contribute most to neighborhood crimes rates. This is a problem in unsupervised learning that we have thought about but will look to implement in the future. There are several approaches to consider in dimensionality reduction in the feature space, which we have not touched upon yet (except in constructing the Lasso model). Using principal component analysis (PCA) or a greedy step-wise algorithm will help us understand which inputs are truly the most relevant predictors for our task.

Note that we are only making use of 2010 crimes and demographic data in our study because of the completeness of the 2010 Census data. With the understanding that demographic estimates from the American Community Survey for subsequent years have larger variability, we would like to understand how changing demographic factors can influence crimes rates in various neighborhoods. For instance, in areas of high gentrification over the past 6 years, have crimes rates changed drastically as the population has changed? Given a subset of very relevant features from our predictive models, will we be able to identify which neighborhoods have become safer or more dangerous over time? This final step will be the one that will enable us to make predictions for the future and draw meaningful conclusions about the nature and frequency of crimes for specific neighborhoods.