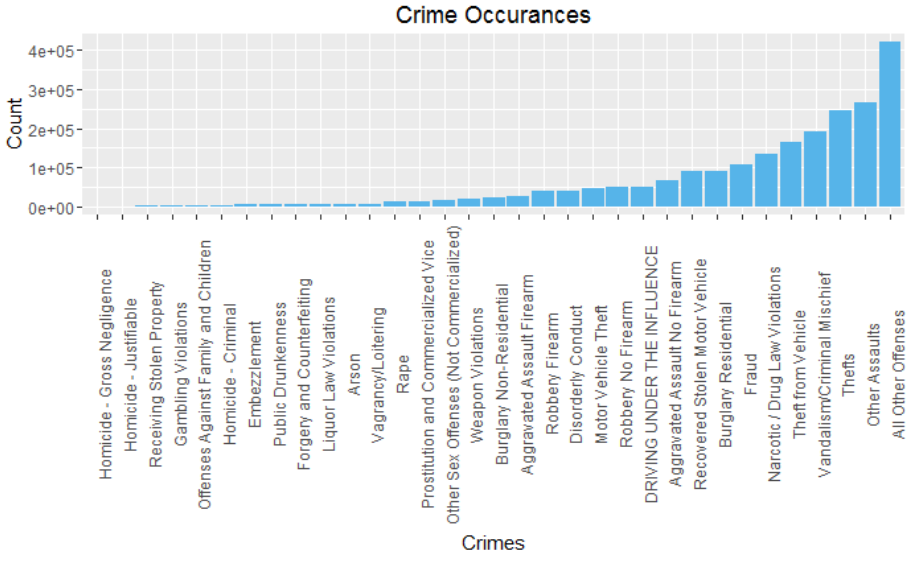
Midterm Report - Predicting Crimes in Philadelphia

Fiona Chin (flc38), Erik Enriquez (eee37), Yumeng Niu (yn227)

October 29, 2016

Description Of Data

Our goal for the project is to predict the number of different crimes in Philadelphia. That way, the police can predict and be prepared for future crime. We plan to reach this goal using Philadelphia Crime Data. This dataset comes from OpenDailyPhilly.org, whose primary goal is to make data available to promote business and reach out to its citizens. This dataset dates from 2006 to 2016 and consists of 14 features and 2,154,123 examples of different crime incidents reported in Philadelphia. The data can be split into time-related information, location-related information, and type of crime related information about the crimes in Philadelphia. But, there wasn't enough features to build a crime predicting model in the main dataset. Thus, the team decided to focus on how crime data is affected by weather patterns and yearly general statistic indicators. The Philadelphia weather dataset we found from weather.gov dates from 2006 to 2015 containing indicators such as temperature, precipitation and snow depth (3652 examples, 10 features). The yearly statistic indicators were extracted from OpenDailyPhilly.org from the Community Health Assessment containing yearly data from 2000 to 2012 (15 examples, 8 features). It consists of unemployment rates, cancer mortality, homicide mortality, and other indicators over the years. Overall, the main dataset provides a variety of crimes to analyze and predict on. In order to predict these crimes over time, we will have to clean the data to look for the most important features to model on.



Data Cleaning, Reshaping, and Joining

Cleaning

Since only a small of percentage of our original Philadelphia crime dataset had missing values (0.08%), we decided to get rid of those rows. The rest of the data appeared to be in good condition since there were no significant outliers in the dataset.

Reshaping

Because we are interested in predicting number of criminal incidents, we had to aggregate and pivot data so that our y values are sum of incidents. We chose to only use the month (e.g. 2006-01) and type of crime features from our original Philadelphia crime dataset to pivot on. We discarded other features because we did not want to segment our data into too fine a granularity, which would results in very small number of incidents in each bin. For a similar purpose, we decide to aggregate similar crimes into a single category. For example, we grouped "Burglary Residential" and "Burglary Non-Residential" into a single category called "Burglary". With these measures, we reduced the number of different types of crime from 33 to 25.

Feature Engineering

Since type of crime is categorical, in order to put it in our X-matrix, we transformed each type into a column. For example, the "Burglary" column would only contain a one for a row if that row corresponds to that type of crime. In all other cases, it would contain a zero.

Joining

We joined that Philadelphia crime dataset with a dataset on Philadelphia monthly weather statistics (i.e average temperature, precipitation and snowfall) and with a dataset composed of monthly general statistics specific to Philadelphia (i.e unemployment rate, suicide mortality rate, etc). We joined them on Year and Month respectively. (We transformed the raw weather dataset into monthly data by calculating the mean of daily values for each month. This measure is performed to be consistent with the granularity of our primary dataset — Philadelphia Crime).

Additional Cleaning

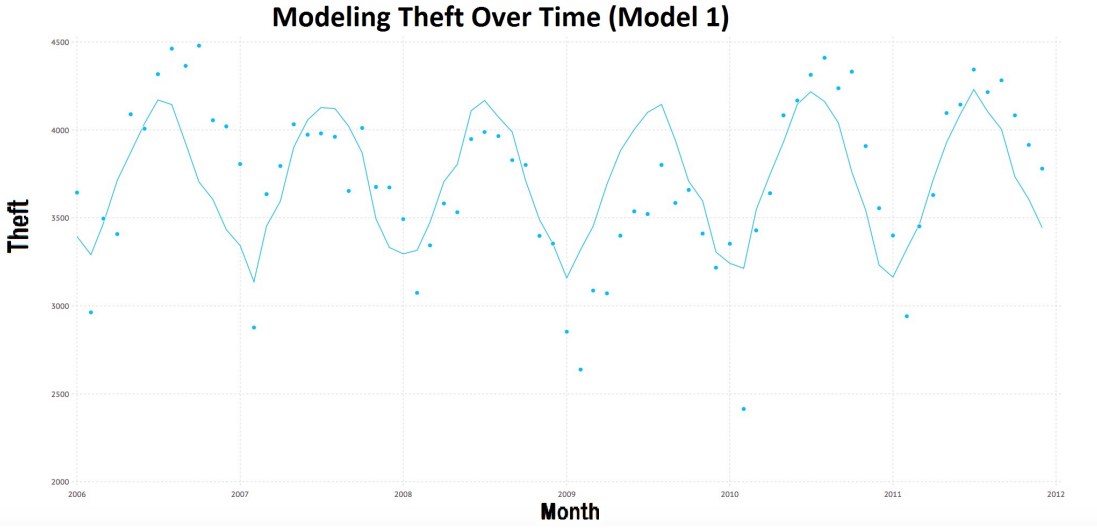
Because the dataset containing general Philadelphia statistics only contained data up to 2012, we decided to only use 2006-2012 data.

Training Set vs Test Set

Because our data is a time series, we decided to use data corresponding to the years 2006-2011 as training data, and 2012 data as test data. Our data set now consists of 44 features and 2,099 examples.

Preliminary Model Fitting

As we mentioned before, our cleaned dataset contains 25 unique types of crime. To start with, we want to experiment with a single type of crime, and gather some additional insights in potential models we could employ. Therefore, we want to first just look at Thefts, the most common type of crime. For these records, we will also start with a rather simple model. y = [Number of Thefts Incidents For Each Month]. It seems from the graph we plotted for all type of crime that there seems to be annual seasonal trend. Hence we started with plotting an individual plot for Thefts as below:



This graph confirms our insight that there exist some seasonal trend fluctuating similarly each year. The first thought we had was setting X be the month number (1 for Jan, 12 for Dec etc.). However, the correlation between month number (increasing from Jan to Dec) and number of incidents (seasonally fluctuating up and down) does not appear to be linear or polynomial. Thus, the month number is not a good X feature. However, if we plot the number of monthly Thefts incident against average temperature, there seems to be a linear correlation.

Model 1: X = [Offset; Average Temperature of the Month] (Least Square Loss)

From the in sample fitting plot, temperature seems like a good indicator for the number of incidents in this month. However, it seems like the model predicted 2009 criminal rate far above actual. Therefore, we also want to examine if we could add another predictor to help with this problem. Hence, we have the second model: adding feature number of incidents 12 months

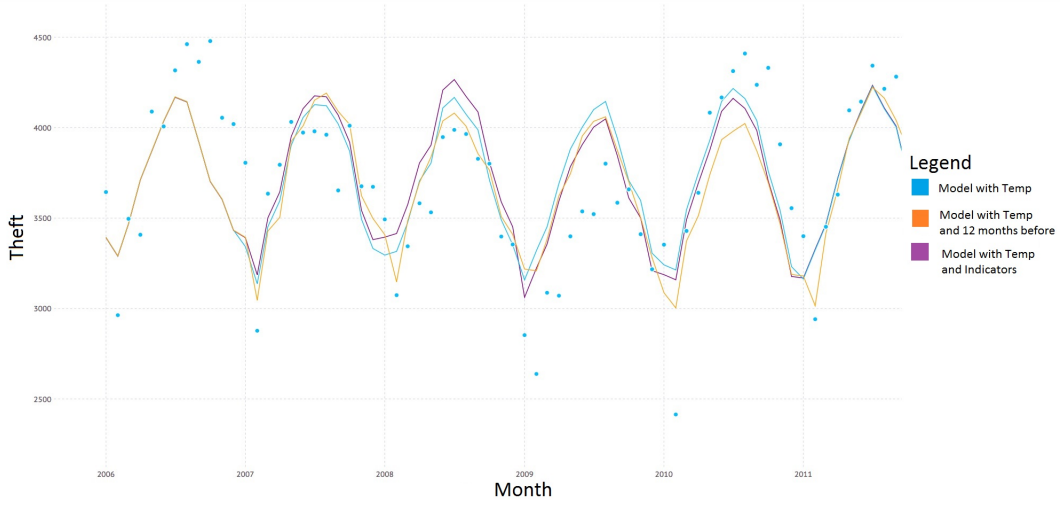
before (e.g. 2015-01 for 2016-01). Looking at a scatter plot between this new predictor and our y value, there seems to be some positive linear correlation.

Model 2: X = [Offset Average; Temperature of the Month; Data 12 Months Before]

It seems like the updated model with Temperature and previous 12 month data actually did even worse for 2010, due to the dragging (low rate) for 2009. However, if we look at error on training set, the updated model has a small in sample error (which make sense because we are fitting in more information). Now, we also want to see if any of the general statistic indicators have a big impact on our model. From a brief look at the annual indicators, it seems like unemployment rate for 2009 is especially low, and number of day of good air is especially high for 2009. We wonder if they could explain the drop in crime rate that year.

Model 3: X = [Offset Average; Temperature of the Month; Unemployment; Number of Days with Good Air Quality]

Interestingly, this new model does not seem to do a lot better than the first model just from looking at the graph.In addition, the square error is actually higher for than the second model, and not remarkably smaller than the first model:



Future Plans

Over (and Under-) fitting:

We plan to avoid underfitting by trying different models with different features to see which features are the most important. We have already reduced overfitting by decreasing the number of features in our data. In addition, we plan to avoid overfitting by adding regularizers such as LASSO to produce sparser solutions. Sparsity will allow for better representation of our large data set. Finally, we plan to implement k-cross validation to prevent overfitting.

Testing Effectiveness:

We will use our test set when we have decided the final model. Namely, we will look at the out of sample error as an estimation of the effectiveness of the model.

To Dos:

We are planning to explore more possible models for the Thefts dataset. Some possible ones are additional weather statistics, additional statistical indicators. We also want to see if we could modify our dataset to deseasonalize it by dividing every y with the mean of this month (over all years), and then use previous number of incidents in last month (instead of last year) as a feature. In addition, we could explore other loss functions or regularizers. Although our y values are sum of incidents per period, which would be natural to approximate by normal distribution. With these additional models and feature engineering approaches, we need to analyze the bias and variance of each model to get more insight into which model to use. Lastly, we want to check how well our model generalizes. Thefts is just one type of crimes we have in our dataset. We have to decide if we want to fit one model for all of the types. (Either use one column for one type of crime in X or aggregate all the type in same month into one sample), or to fit an individual model for each type of crime.)