**Predicting Crimes in Philadelphia**

### 

### Hong Son, Raj Patel, Kunal Sharma

### Drexel University, 2020-2021 Winter Quarter

### DSCI592 Data Science Capstone II

**Identifying Dataset(s)**

**Philadelphia Crime Data:**

<https://www.kaggle.com/mchirico/philadelphiacrimedata>

For our project, we would like to conduct an analysis on a Philadelphia crime dataset published by OpenPhillyData. The dataset dates from 12-31-05 to 03-22-17 (majority of the dates are from 2006-2016). There are 14 total columns in this dataset. They include the district number, sector, dispatch date, dispatch time, location, general code, descriptions of the crime, and the district. This dataset was published to the public so people can assess the crimes happening in Philadelphia.

**Weather Data:**

<http://www.climate.psu.edu/data/city_information/index.php?city=phl&page=dwa&type=big7>

To assist in getting more out of our analysis, we will be scraping Philadelphia weather data for Philadelphia from 2007 to 2018. The dataset includes the temperatures in Fahrenheit, the max and min temperature, environment information, wind, and precipitation. The data was formatted into PDF files and a dataset was formed [~4,000 rows and 15 columns].

**Parking Violations Data:**

<https://www.opendataphilly.org/dataset/parking-violations>

We will be using the latitude and longitude points to map the zip codes in this dataset to the crime dataset.

**Understanding the Target Data and Substantiating its Existence**

Potential users of this analysis are law enforcement agencies so that they can dedicate resources and assistance to prevent recurring crimes. Our proposed solution is to create multiple models that can predict and classify the type of crime. This can also help provide community insights on how to make improvements to reduce the crime from occuring in the first place. Another solution we intend to work on is predicting which neighborhoods are likely to increase and decrease in crime rates.

**Data Cleansing/Transformation**

Philadelphia Crime/Weather Data:

* From last term, these two datasets were cleansed and integrated based on one of the common attributes, zip code. We will now use this transformed data set as our new “base” dataset.

Philadelphia Housing Data:

* Data cleansing would include removing of rows with null property values to have zip-codes in the dataset which matched the crime data set.

**Machine Learning:**

The machine learning part of the project will be made up of three components: data transformation/normalization, regression, and classification. For classification, we will do our best to implement all of the following model types: Logistic Regression, Naive Bayes, SVM, Decision Tree Classifier, Randomforest, Neural Networks, LDA, QDA, KNN, NN. The next part is regression, which will consist of all of the following: Linear Regression, add regularizer to linear regression (Ridge and Lasso Regression), Decision Tree Regression, Support Vector Regression, Xgboost, Neural Networks (NN), partial least square regression, In addition, for data transformation and normalization, we will perform at least one of the following: Dimensionality reduction (PCA as an example), Hyper-parameter tuning (for example gamma and C for SVM), and Feature Selection.

**INTENDED STEPS**

The schedule column indicates the week we will work on the task.

|  |  |  |  |
| --- | --- | --- | --- |
| **Schedule** | **Task** | **Ownership** | **Status** |
| Week 1 | Weekly Team meeting to Discuss Project | Team | DONE |
| Week 2 | Data science capstone project launch report (Use DSCI 591 template) | Team | DONE |
| Week 3 | Outline and steps to accomplish the capstone, Identifying Data Sources project | Raj | IN PROGRESS |
| Week 4 | Start coding in Python and/or R | Team | IN PROGRESS |
| Week 5 | Pitch Presentation | Raj | IN PROGRESS |
| Week 6 | Data Acquisition, Pre-processing and EDA report (Use DSCI 591 template) | Hong | IN PROGRESS |
| Week 7 | Applying Machine Learning Models | Kunal | IN PROGRESS |
| Week 8 | Model Evaluation and make the model better  - Data transformation/Normalization  - Dimensionality reduction (PCA as an example)  - Hyper-parameter tuning (for example gamma and C for SVM)  - Feature Selection |  |  |
| Week 9 | Predictive modeling report (Find its template on Blackboard) |  |  |
| Week 10 | Final Project Presentation |  |  |