

Predicting Crime in Boston

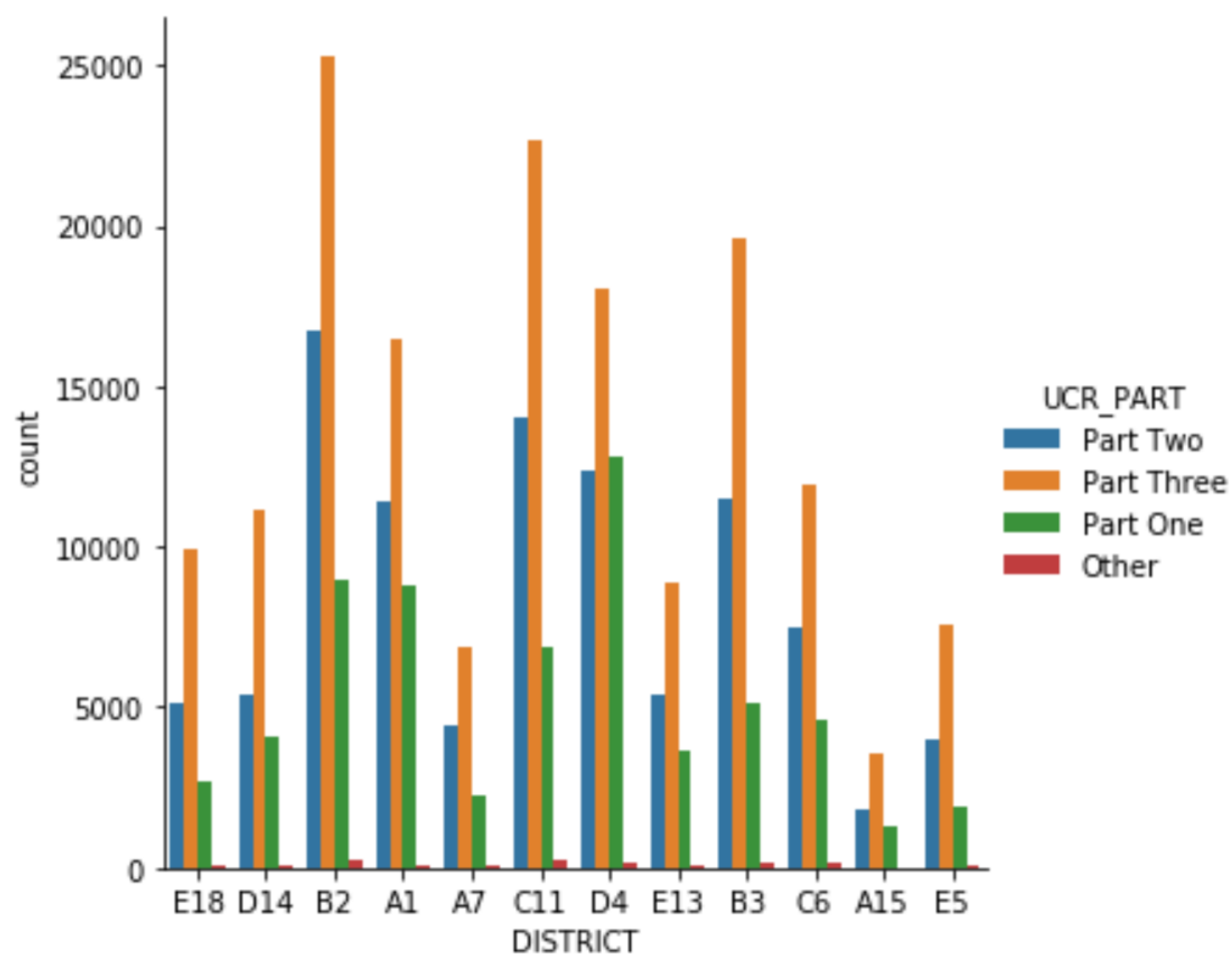
Chris McCooey, Julie Osborne, Mary Wishart

Abstract

Boston is the largest city in New England. Being a major metropolitan area means that crimes occur around the city every day. Our project takes a look into the “Crimes in Boston” dataset. This is a dataset of police reports from 2015-2018. From the information in these data, we used machine learning techniques to see if, given a specific crime and the day which that crime occurred, we could predict where that crime is most likely to happen within the various neighborhoods in Boston.

1 Introduction

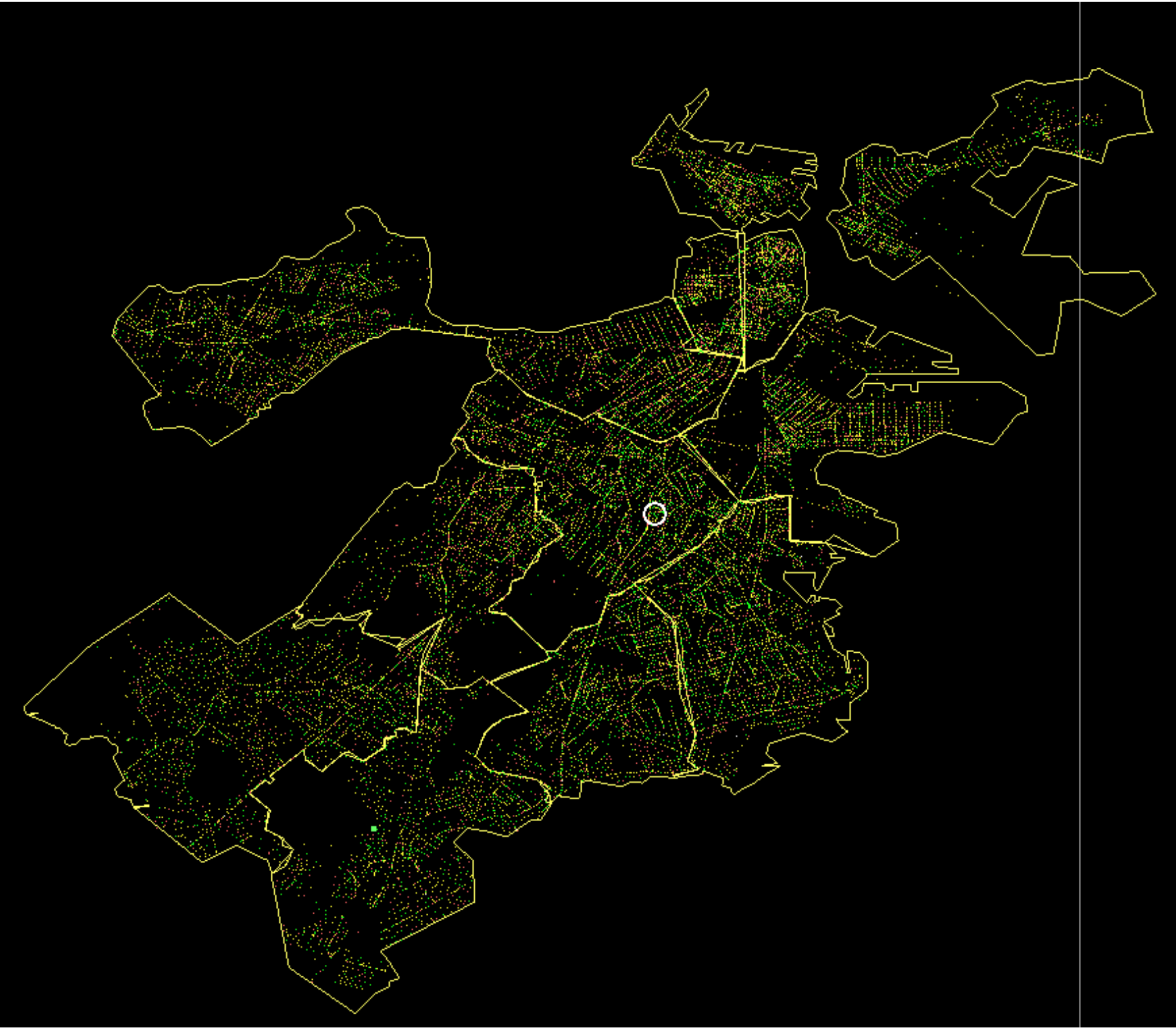
- Purpose: Try to predict where various crimes are likely to happen within the city of Boston
- Use Multi-class logistic regression to predict where a specific crime was likely to happen
- Use Decision Tree algorithm to predict where a crime was likely to happen
- Create a map of Boston which visualizes our results



A graph showing the police districts along with the amount of crime organized by UCR part number.

2 Setup to our model

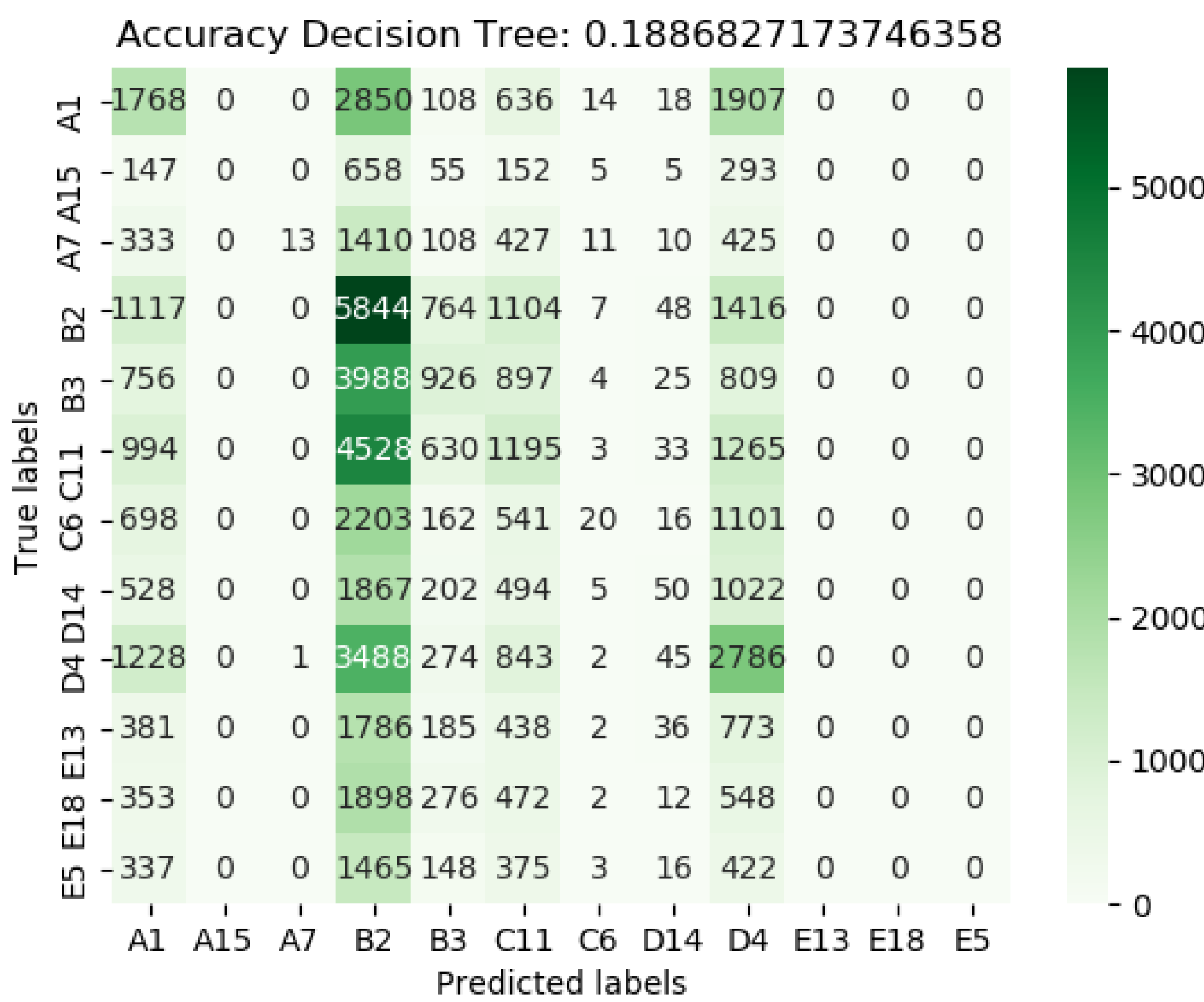
- Each location of city is organized by BPD districts
- 12 districts in total
- Individual crimes are given an offense code number. These are then put into offense groups
- For our analysis, we used the offense group, not the individual offense code number
- 67 offense groups in total
- Combined the month and year columns to be one column
- One Hot encoded all of our variables as they were categorical
- Data was spilt into 80/20 training, testing datasets



This map is a visual representation of our dataset. It shows the city of Boston, divided into each police district. Each dot is a row in the dataset. For example, the dot circled represents a police report where the officer was called for medical assistance within district B2 on a Friday in August, 2018 at 10pm. Each location of the dot is determined by the latitude and longitude data. This map is an inactive map using OpenGL and Qt.

3 Decision Tree

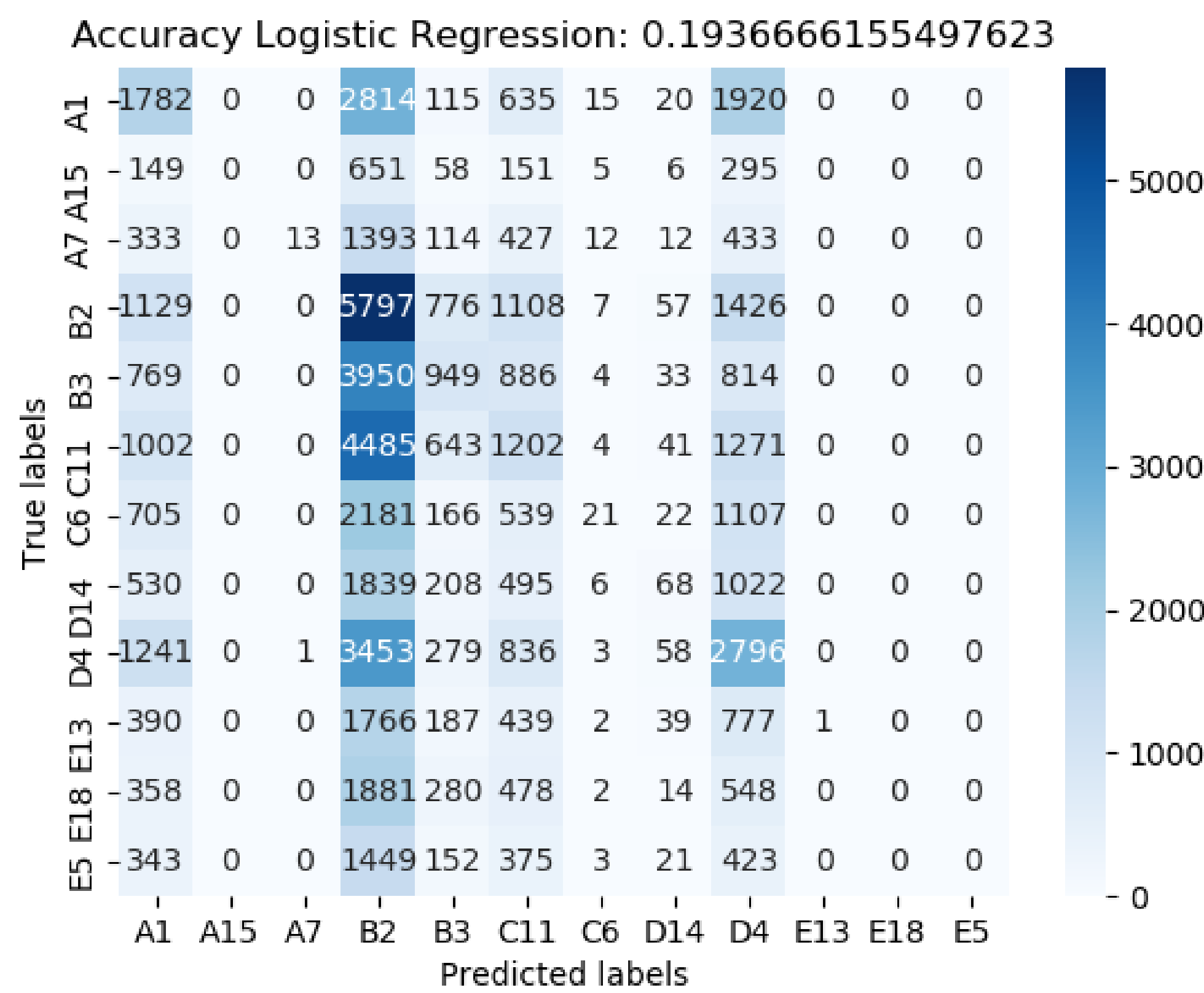
To set up a baseline for our data predictions, we first created a decision tree. To create this tree we used the DecisionTreeClassifier from sklearn. In order to get the maximum depth we ran the algorithm, testing depths from 1 to 119. A depth of 26 gave the highest accuracy. The accuracy for this model was low at only 0.18. This model misclassified a lot of the districts, for example it did not classify any of the A15 crimes correctly and classified a high amount of crimes as B2 when they should not have been.



A confusion matrix representing our decision tree model

4 Multi-class logistic regression

Next we created a multi-class logistic regression algorithm to try to compute where a crime was most likely to happen. To calculate the model, we used sklearn’s built in LogisticRegression method. This model had very low accuracy of about 20%



A Confusion Matrix representing our mutli-class logistic model

5 Future Work

- Increase prediction accuracy of our models
- Bring other data into consideration for classification such as time crime occurred
- Analyze other datasets for information on why certain districts have more crimes or are more likely to have specific crimes than other districts
- Create a visualization of our data using the map of Boston. Each dot on the map would be a prediction where the color of the dot would represent whether or not the prediction was correct.

Citations

Jain, A. (2018, February). Crimes in Boston, Version 3. Retrieved November, 2019 from <https://www.kaggle.com/ankkur13/boston-crime-data>.
Boston Police. “The Boston Police Department’s Virtual Community.” Bpdnews.com, 7 Dec. 2019, <https://bpdnews.com/>.