Assignment 1: Project Goals

MSDS 498 (Spring 2020): Bryan Bruno, Michael Haugan, Ksenia Luu

Data Collection Plan

We are planning to utilize three different sources for local crime report data in Dallas, San Francisco and Minneapolis. Each dataset has a slightly different set of variables and crime classifications. Along with the main crime incident data, we want to enrich our analysis by utilizing outside sources. We are hoping to find data sources that can be used across all three locations in order to conduct a final analysis comparing all three models. The important piece to consider is the level of time granularity we want to use as this will heavily dictate the type and quality of external data we can use. Ultimately, we have to weigh the tradeoff of being able to make more specific time and location predictions versus the open source availability of the external data we want to potentially bring in.

San Francisco - Bryan

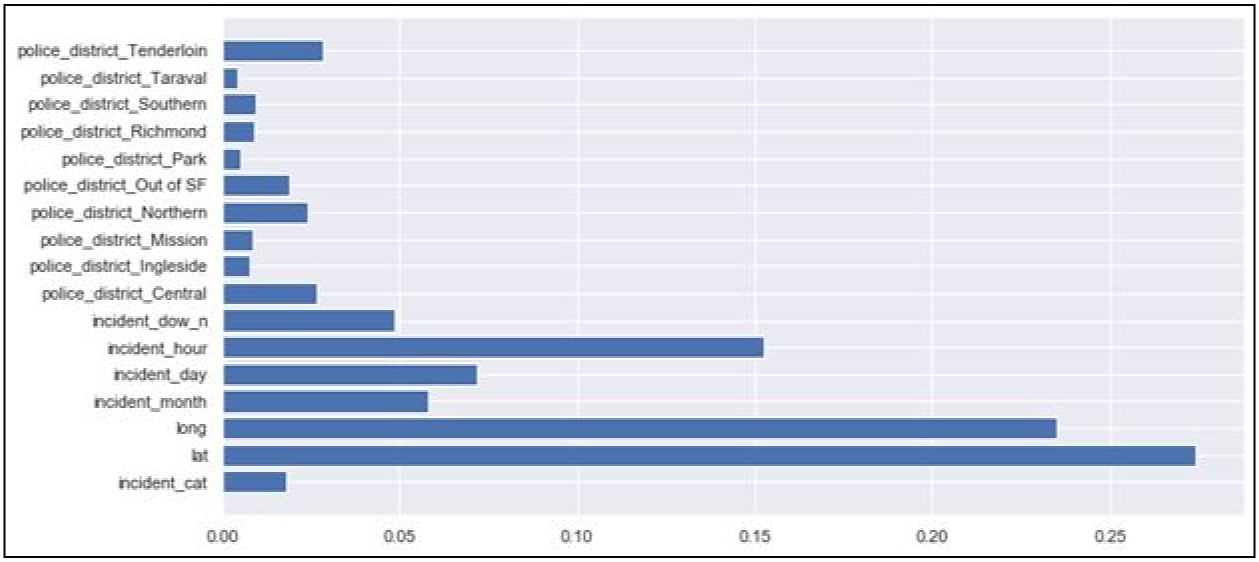
The data selected for this regional exploratory data analysis is from the publicly available source of the San Francisco Police Department's website. San Francisco is a small but highly dense city in terms of population, visitors and workers. In the past two years (2018 and 2019), this city has had over 300,000 incident reports. Upon preliminary data cleansing procedures, this has been reduced to 285,509 viable incidents for exploratory data analysis and machine learning purposes.

From this particular dataset, the key features that have remained are as follows:

- Incident Category
- Longitude & Latitude
- Hour, Day, Month, Year
- Police District

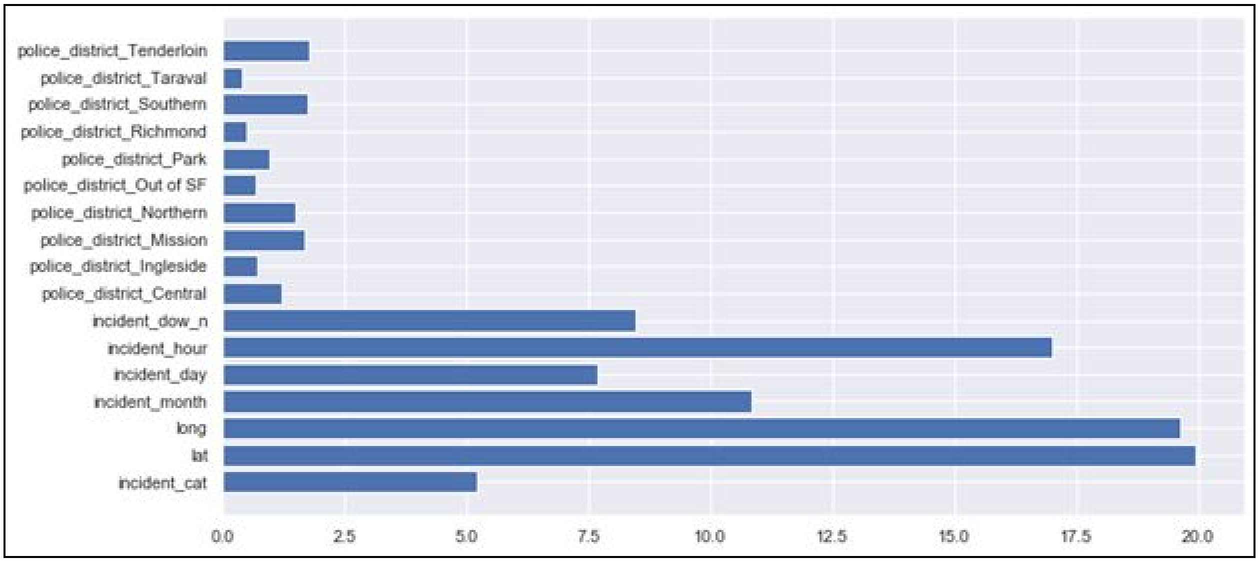
All other features are either redundant or specific to the San Francisco Police Department, which cannot be used for predictive modeling. This data source does not have enough features to properly train a model and satisfy the requirements of this project. As such, bringing additional data has become a priority. Figure 1 and 2 below show the importance of the aforementioned features.

Figure 1



Note. Feature importance using Random Forest Classifier.

Figure 2



Note. Feature importance using CatBoost Classifier.

Longitude, Latitude, and Incident Hour are the three most essential features for model generation in both visualizations. Using this information, additional location and hourly based time data should be introduced into the existing data. Two possible sources are:

- Weather data that is tracked by longitude and latitude along with hourly updates.
- Stock market/economic data as another time related source.

Minneapolis - Michael Haugan

I am basing my exploratory analysis and research off the city in which I live, Minneapois. I am utilizing the open source Minneapolis Police Incident data provided by Minneapolis Open Data. This dataset is based on all crimes reported (where a crime report was filed) by the Minneapolis Police Department. My initial research is based on the past 10 years of crime data from January 1st, 2010 through December 31st 2019. In total it is slightly over 213K incidents with about 20 variables within each dataset. One thing to note as we each bring our research and analysis together is the historical time period we want to use to train our predictions on. Here, I am using 10 years. Ksenia is using about 5 and Bryan is similar as well. Because of the sheer size of data this could lead to, especially when considering the potentially hourly data we are going to bring in externally, I may need to shorten my historical period down to something like 5 years as well. In either case, we will ideally want each of historical periods to be the same to enable easier comparison across models and predictions.

The data elements of most importance from the Minneapolis crime data are:

- Timestamp of incident (in Central Standard Time)
- Geographic location of incident (based on latitude and longitude)
- Neighborhood incident occurred in
- Precinct incident was allocated to
- Offense committed
- Short description of offense

These elements give me great historical understanding of where, when and what incident occurred. I am able to learn a ton just from this. Like, what areas experience the most crimes? What crimes are most prevalent in said areas? What times during the day do most crimes occur? However, there is a ton of opportunity to enrich this data even more by bringing in more sources of information.

For starters, I have utilized historical Minneapolis weather data at the hourly level from openweathermap.org to join into my current incident dataset. In Minnesota, we have very high extremes when it comes to weather. It gets very hot in the summer (with high humidity at times) and also very cold in the winter (with a lot of snow at times). Bringing this information into my crime dataset allows me to see how much of an effect weather has on crime. Does crime spike in the summer when the weather is nice? Does crime go away in the winter when it is absolutely freezing outside? Does crime fluctuate based on the amount of precipitation in the air (either rain or snow)? Through this exploratory phase I can’t completely determine if there is just a correlation between weather and crime or a true causation, but that is what the continued research and eventual model building process is for.

Dallas - Ksenia Luu

I will be utilizing the public Dallas Police Incident data provided by Dallas Open Data. The dataset reflects daily crimes reported by the Dallas Police Department consisting of approximately 590,000 incidents from June 1, 2014 to December 31, 2019. The dataset is filtered to exclude potentially sensitive and confidential information prior to being made available to the public. Dataset includes 100 variables from which I will be utilizing the following set to conduct initial EDA and model development:

- Date and Time of incident
- Location of incident (latitude/longitude)
- Incident Type - a new variable created in order to unify 60% of dataset records that are tagged with National Incident-Based Reporting System (NIBRS) codes and the other 40% of records that are tagged with Uniform Crime Reporting (URC) codes
- UCR Offense Type - Offense category Part1 or Part2 or Not coded
- NIBRS Group -
- Council District Division - Geographic area comprised of city council districts where incident occurred
- Zipcode - Zipcode in which incident occurred

Dallas Police Stations - location data for eight police stations across the City of Dallas
Dallas Fire Stations - location data for 57 fire stations across the City of Dallas

Other Potential Data Sources:

- Weather - hourly temperature and weather conditions by zipcode/city from Open Weather
- Housing Price and Rental Rates - housing inventory, housing sales, rentals, and home value data by US metro area provided by the Zillow Research Team.
- Demographic Data - individuals and household data from Decennial Census and other surveys administered by the United States Census Bureau
    - Annual Social and Economic Data (Current Population Survey)

Lastly, we do want to mention that further information will continue to be brought into this exploration as we as a group continue exploring the potential availability and viability of external data sources. There is definitely more information that we can feed into a model. We just need to figure out one, the right information to bring in and two, how readily available is this data? Some areas of thought for further data to include are; socio-economic data, housing/rental data and even major event data that includes concerts, sporting, political, etc.

Team Responsibilities

Each group member will dedicate their time to perform our respective exploratory data analyses. We communicate regularly through Slack and Zoom calls to discuss our findings to better improve our understanding of the information. At the beginning stages, we claim responsibility for Dallas, San Francisco and Minneapolis. However, we will eventually bring our results together to expand the scope of our project. By cross referencing each other's results, we will test for a larger and more generalized model. If this does not work as planned, we will keep our models specific to their respective cities and use each other's results for comparison purposes.

Desired Outcomes

Our desired outcome by the end of this ten-week course is to develop models that will be accurate in predicting the time and location of violent crimes in Dallas, San Francisco and Minneapolis. The long-term goal for this project is to productionize our models and develop an application for the local Law Enforcement enabling them to optimize planning and resource allocation.

Final Project Portfolio

This project will be showcased in a dedicated repository on Ksenia's personal GitHub page. The repository will include a project README outlining the project objective, data sources, and repository contents including the following:

- Final Report - pdf of the final team report
- Presentation materials - slides, audio, video, etc.
- Exploratory Data Analysis - Jupyter notebook with dataset EDA
- Model - Jupyter notebook showing the development and testing of the final ML model

The link to the Github repository will be added to Ksenia's Linkedin page and Resume. Bryan and Michael will clone the repo into our own personal Github page's as well.

Michael has also published an article on Medium.com that outlines the problem statement and initial EDA we are going through. He will also publish an additional article once we finish the project outlining our findings, model approach and predictive accuracy. Publishing articles has been a great way to expand the professional portfolio in a more public and visible way.

The end product will be something that can not only showcase the tremendous work we put into it within our professional portfolios, but also a product that potentially could be sold to interested Law Enforcement agencies. All of which will show a wide range of skills across domain research, external data source investigation and inclusion, data pre-processing, model training and validating and finally wrapping it all up within a business context to showcase operational value for the customer (in this case the PD’s).

Optional/Undetermined:
Present findings to IBM and potentially have our work published and patented through them.

References

Astral v2.1 (n.d.). Retrieved April 26, 2020, from https://astral.readthedocs.io/en/stable/index.html

Alphabet Inc. Class C Capital Stock (GOOG) Historical Data. (n.d.). Retrieved April 25, 2020, from https://www.nasdaq.com/market-activity/stocks/goog/historical

Dallas Police Department. (n.d.). Police Incidents: Dallas OpenData. Retrieved April 6, 2020, from https://www.dallasopendata.com/Public-Safety/Police-Incidents/qv6i-rri7

Dallas Police Stations: Dallas OpenData. (2013, September 30). Retrieved April 20, 2020, from https://www.dallasopendata.com/Public-Safety/Dallas-Police-Stations/si5f-a98p

Fire-Rescue, D. (2013, September 23). City of Dallas Fire Station Locations: Dallas OpenData. Retrieved April 20, 2020, from https://www.dallasopendata.com/Public-Safety/City-of-Dallas-Fire-Station-Locations/w75g-c54n

Libraries. (2013, August 14). City of Dallas Library Locations: Dallas OpenData. Retrieved April 22, 2020, from https://www.dallasopendata.com/City-Services/City-of-Dallas-Library-Locations/2ksy-mdcf

Open Minneapolis. (n.d.). Retrieved March 28th, 2020, from http://opendata.minneapolismn.gov/search?groupIds=79606f50581f4a33b14a19e61c4891f7

OpenWeatherMap.org. (n.d.). C urrent weather and forecast. Retrieved April 2, 2020, from https://openweathermap.org

San Francisco Open Data. (n.d.). Retrieved April 26, 2020, from https://data.sfgov.org/

URC Data Online. (n.d.). Retrieved April 15, 2020, from https://www.ucrdatatool.gov/twomeasures.cfm

U.S. Department of Justice—Federal Bureau of Investigation. (n.d.). Retrieved April 25, 2020, from https://ucr.fbi.gov/nibrs/2012/resources/a-guide-to-understanding-nibrs

US Census Bureau. (2020, February 27). Demographic Data. Retrieved April 26, 2020, from https://www.census.gov/programs-surveys/ces/data/restricted-use-data/demographic-data.html

Housing Data. (n.d.). Retrieved April 26, 2020, from https://www.zillow.com/research/data/