Capstone Project Proposal

San Francisco Crime Classification

Anushree Srinivas

Mentor: Srdjan Santic

Problem to be solved and Motivation:

From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz.Today, the city is known more for its tech scene than its criminal past. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay.

From Sunset to SOMA, and Marina to Excelsior, this competition's dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. The objective of this project is to predict the category of crime that occurred at a given time and location.

Client:

City and County of San Francisco is trying to make the city and the neighborhoods more safe. The City authorities will be the client for whom the conducted data analysis as part of the capstone project will be beneficial.

Data:

This dataset is obtained from SF OpenData, the central clearinghouse for data published by the City and County of San Francisco which contains incidents derived from SFPD Crime Incident Reporting system. It provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. The date ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7... belong to test set, week 2,4,6,8 belong to training set.
The data fields include, dates, category, descript, DayOfWeek, PdDistrict, Resolution, Address, X(Latitude), Y(Longitude). The data will be obtained from a previous Kaggle Competition.

https://www.kaggle.com/c/sf-crime/data

Approach:

Initially, I am going to dive deep into the data to get a thorough understanding of the different variables. I plan to do some data wrangling and cleaning to deal with missing values etc. Once the data is cleaned, I am going to perform some exploratory analysis to recover some interesting insights from the data. Feature engineering will also be applied to discover some new meaningful features that can be used while building the models. Finally, different models will be built using machine learning algorithms. Data visualizations will help in communicating these insights and tell a story.

Deliverables:

The capstone project will produce a Presentation (slide deck), a report (highlighting the approach, findings etc.) and the Python Code in the form of a iPython Notebook that will be submitted into my Github repository. A final paper explaining the problem approach, findings and recommendations will also be submitted.