Project: Crime in San Francisco

Final Report

IEOR 135 | Spring 2020

Nhut Nguyen
Chad Wakamiya
Anna Stepnova

May 8, 2020
GitHub link: https://github.com/nhut-nguyen1503/IEOR_135_Project

Problem

The issue of rising criminal activity in the city of San Francisco is truly alarming. Not only it is worsening every year, but the city is a nation's leader in crimes including burglary, larceny, shoplifting, and vandalism. The ability to effectively combat the current situation is significantly reliant on effective distribution of human resources in San Francisco police department. Needless to mention, that real estate investors and casual home buyers have question of current and prospective safety high on their list of location requirements. The granularity of data currently available is not easily interpretable and useful. We plan to develop a simple and effective dashboard that allows a potential user to quickly access the machine learning enhanced forecast to their desired level of geographical granularity.

Data

We collected public records of all crimes occurred in San Francisco from 2003 until 2018 that we found on *datasf.org*. We cleaned the data by removing observations with NA values and optimizing data for uniformity. We converted temporal and geographical data into useful formats in preparation for data evaluation. We ended up with 37 features among which are category, description, day of the week, day, time, police district, address, and coordinates of a crime.

Data Exploration

We've performed several visualizations on cleaned and formatted data. We explored top 10 common crimes in San Francisco and observed high disparity in their frequencies. Further we examined density of distribution of the crime per city district. We plotted the distribution of each crime type through the available timeline and observed no common or steady trend among the majority of them. We looked at boxplot crime distributions per year and per month (looking for seasonality). The boxplots confirmed the steady rising trend from the year of 2014 and on.

Models

For this project, we extensively investigated several different approaches. In the end, we have narrowed our models down to six useful models:

- Seasonal Auto Regressive Integrated Moving Average (SARIMA)
- Triple Exponential Smoothing
- Random Forest
- Adaboost Regressor
- Gradient Boosting Regressor
- Ensembling

We chose our Baseline Model to always predicts the average crimes per month.
The following error metrics were used to compare models:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

Here are the values for the above metrics for all our models for out of sample evaluations:

Evaluation **Model**	MAE	RMSE	MAPE
Baseline Model	*95.2*	*110.92*	*8.13%*
SARIMA	51.96	60.53	4.57%
Exponential Smoothing	58.28	65.05	5.14%
Random Forest	67.93	84.29	5.92%
Adaboost	62.57	78.59	5.42%
Gradiend Boosting	67.54	85.93	5.84%
Ensembling Model	42.19	60.52	3.69%

As seen above, all of the models improved compare to the Baseline model. Before deciding on Ensembling Model we observed that some models generally overestimate while other underestimate (based on the graphic output of forecasts). As a result of the observation we decided to proceed with Ensembling Model for final forecasting that improved the overall performance.

Result

We used Voila package for the UI. The final product is a simple and effective dashboard that allows an end user to select a district or enter a physical address and the crime of the interest to have a predicted trend to be outputted on our interactive dashboard. The dashboard has an option to over plot forecasts in two locations for graphic output that is very helpful for a visual comparative evaluation.

Impact and Future

Systems reinforcing educated decisions are on high demand in many fields. We believe we've delivered a potent tool to increase confidence in making essential personal and financial decisions more safely. We also believe that the tool may enhance the quality of SFPD performance.

We are convinced that future applications of the system are possible in fulfilling the needs of other cities and different industries. For convenience of personal use, the tool can be also realized in the form of mobile app for easy access and on-the-go consultation.