

Machine Learning Engineer Nanodegree

Capstone Proposal

Ken Adachi

July 2nd, 2019

Proposal

Domain Background

I want to apply Machine Learning technique for the purpose of crime prevention. Crime Opportunity Theory suggests that the occurrence of a crime depends not only on the presence of the motivated offender but also on the conditions of the environment in which that offender is situated.

reference: https://en.wikipedia.org/wiki/Crime_opportunity_theory

I expect that I can find a pattern in where actual crime happened. So I want to analyze the historical crime incident data with the information on where it occurred.

This might lead to prevent crime to happen in my neighborhood and protect children in my community.

Problem Statement

In order to avoid to be involved in encountering crime, I want to predict if the crime occurs in the present location.

Datasets and Inputs

I'm going to use "Police Department Incident Reports: Historical 2003 to May 2018" data set provided by San Francisco City Government. <https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry>

This dataset includes police incident reports filed by officers and by individuals through self-service online reporting for non-emergency cases through 2003 to 2018. The dataset has attributes such as when the incident reports were filed (Date, time) and detail location of the incidents (latitude, longitude).

Optionally, I want to use the data provided by Japanese Local Police.

https://www.keishicho.metro.tokyo.jp/about_mpd/jokyo_tokei/jokyo/ninchikensu.html However, this dataset is a summary data and lacks the detailed information such as when the incident happened or the exact location. This would not lead to solution design.

Solution Statement

I want to build prediction model leveraging machine learning techniques.

In this case, I want to apply supervised learning because the labeled data is available.

Based on sklearn algorithm cheat-sheet, I'm going to use classification methods to build the model.

https://scikit-learn.org/stable/tutorial/machine_learning_map/

Benchmark Model

Found similar analysis below. The model received 61% accuracy on evaluation.

<https://medium.com/@m.vkumar89/san-francisco-spatial-data-research-for-crime-classification-1a6f1c1b7d09>.

Evaluation Metrics

The model should be evaluated using Accuracy score, F measures.

Project Design

Summary of the project design is as follows:

- Problem definition I want to predict the possibility if I'm going to encounter with crime in the current location. This will lead to prevent to encounter with crime and save the people of the local community.
- Choose the algorithm I want to use supervised learning methods, and classification algorithm to build a prediction model.
- Analyze the data Before going into the detail of machine learning, I will conduct data analytics or visualization to find when and where the crime occurred.
- Pre Process Data set has a text field of 'Category' and 'Description'. Although I have not decided yet to use those features for prediction, but when I use it, I need to pre process by tokenizing those text data.
- Train and Tune Divide the data set into training data set and test dataset. Train the model using training data set and then evaluate it with test dataset. Tune hyper parameter to improve the score of the model.