SAN FRANCISCO CRIME CLASSIFICATION

Udacity - Machine Learning Nanodgree Capstone Proposal

By:

Yousef Mohamed Zook

Computer and Systems Engineering Student
Alexandria, University

yousefzook@outlook.com

March, 2018

Introduction:

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.

This crime category prediction problem is similar to other problems like predicting the crime rate which is discussed here and here. Also, temperature prediction using machine learning, this problem uses the locations and some features like humidity and sunshine to train the model and make it predicts the right temperature at the right time. This problem is discussed more in this IEEE doc and this paper.

Domain Background:

Supervised learning is one of the most promising field in machine learning and used to achieve many predictions used nowadays even in business goals. Using machine learning techniques to predict the crimes area and category shows the power of technology in the field of safety.

My self-motivation is that I am aiming to make the world better and safer place, the technology provides a great help to achieve this. Also, applying this on San Francisco problem improves the ability of investigating my area of knowledge in crime prediction a lot. Also I chose a Kaggle problem to be more familiar with Kaggle communities and problems.

---

problem statement:

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. This is a multi-class classification problem, given time and location, the goal is to predict the category of crime that occurred.

---

datasets and inputs:

The dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. The data is divided to 2 sets, Training set and Testing set. 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.

Data fields

1. **Dates** - timestamp of the crime incident
2. **Category** - category of the crime incident (only in train.csv). This is the target variable that is going to be predicted.
3. **Descript** - detailed description of the crime incident (only in train.csv)
4. **DayOfWeek** - the day of the week
5. **PdDistrict** - name of the Police Department District
6. **Resolution** - how the crime incident was resolved (only in train.csv)
7. **Address** - the approximate street address of the crime incident
8. **X** - Longitude
9. **Y** – Latitude

The training dataset consists of 878049 samples. The target class has 39 unique class. Also, it is noticed that the data is imbalanced as there some targets that isn't appear a lot – ex. TREA. So we shouldn't use accuracy here due to imbalanced data, but we may use multi-class logarithmic loss or F1-score.

Of course, Training set will be divided into training and validation set as Kaggle test set doesn't include the targets.

Each category has the following numbers in training dataset:

WARRANTS	42214
OTHER OFFENSES	126182
LARCENY/THEFT	174900
VEHICLE THEFT	53781
VANDALISM	44725
NON-CRIMINAL	92304
ROBBERY	23000
ASSAULT	76876
WEAPON LAWS	8555
BURGLARY	36755
SUSPICIOUS OCC	31414
DRUNKENNESS	4280
FORGERY/COUNTERFEITING	10609
DRUG/NARCOTIC	53971
STOLEN PROPERTY	4540
SECONDARY CODES	9985
TRESPASS	7326
MISSING PERSON	25989
FRAUD	16679
KIDNAPPING	2341
RUNAWAY	1946
DRIVING UNDER THE INFLUENCE	2268
SEX OFFENSES FORCIBLE	4388
PROSTITUTION	7484
DISORDERLY CONDUCT	4320
ARSON	1513
FAMILY OFFENSES	491
LIQUOR LAWS	1903
BRIBERY	289
EMBEZZLEMENT	1166
SUICIDE	508
LOITERING	1225
SEX OFFENSES NON-FORCIBLE	148
EXTORTION	256
GAMBLING	146
BAD CHECKS	406
TREA	6
RECOVERED VEHICLE	3138
PORNOGRAPHY/OBSCENE MAT	22
Total	878049

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

solution statement:

This problem is solved using supervised learning algorithms. Using many known algorithms and comparing them together will give the best accuracy.

Some algorithms I intend to use:

1. KNN
2. Decision Trees
3. Ensemble algorithms
4. Neural networks
5. SVM
6. Xgboost

Having the given features, we can tune the parameters of each algorithm and produce the highest probability category of crime for each location.

---

benchmark model:

There are 3 types of results that the model can be compared to:

1. Self-implemented models Score: Test the model with it self using several algorithms as I mentioned above and comparing them with each other. For example, making a simple regression algorithm or naïve bays and compare it to the result.
2. Running presolved models and compare it with my model. For example this 2 solved models:
    a. LiblineaR(score=2.5921)
    b. EDA and classification, score = 2.56180
3. Random Score: We can check if the model makes better than a random prediction or not.
4. Kaggle scores: Comparing the results with Kaggle public scores which is a great indicator for how our model performance is.

---

evaluation metrics:

The model prediction for this problem can be evaluated in several ways, but since this problem in Kaggle is measured using multi-class logarithmic loss, I will use this also to measure my model performance and as I said due to imbalanced data

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

where N is the number of cases in the test set, M is the number of class labels, log is the natural logarithm, yij is 1 if observation i is in class j and 0 otherwise, and pij is the predicted probability that observation i belongs to class j.

---

project design:

The project will be processed in steps:

1. Exploring the data
    a. Data reading
    b. Data visualization – using maps, heat maps and contour plots
2. Data preprocessing
    a. Data redundant elimination
    b. Data dimensionality reduction if needed
3. Models implementation
    a. Implement many algorithms to compare them
4. Models evaluation
    a. Evaluate each algorithm and choose the best
5. Tuning best model
    a. Try to make the best better
6. Conclusion and results

Data sets:

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

References:

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

https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice

http://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html

https://stats.stackexchange.com/questions/113301/multi-class-logarithmic-loss-function-per-class

https://www.kaggle.com/rudikruger/liblinear

https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/

https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

https://www.kaggle.com/c/sf-crime/discussion/16289