**Case Study: NSA Warrantless data**

Author: Thip Rattanavilay

Bellevue University DSC-630

**Overview**

The National Security Agency’s mass surveillance has greatly expanded in the years since September 11, 2001. [Disclosures](https://www.aclu.org/nsa-documents-search) have shown that, until recently, the government regularly tracked the calls of hundreds of millions of Americans. Today, it continues to spy on a vast but unknown number of Americans’ international calls, text messages, web-browsing activities, and emails. Yes, they are collecting data from you as we speak.

The government’s surveillance programs have infiltrated most of the communications technologies we have come to rely on. They are largely enabled by a problematic law passed by Congress — the FISA Amendments Act (FAA), which is set to expire this year — along with Executive Order 12,333, the primary authority invoked by the NSA to conduct surveillance outside of the United States. The Patriot Act has also made it easier for the government to spy on Americans right here at home over the past 15 years. Although the Foreign Intelligence Surveillance Court oversees some of the government’s surveillance activities, it operates in near-total secrecy through one-sided procedures that heavily favor the government.

**Business Understanding**

In 2008, the NSA began collecting what it called “upstream” communications, essentially information that travels across the Internet, separate from what it receives from Internet service providers that filter data to respond to agency requests.

The NSA is supposed to be targeting foreign communications, such as e-mail addresses it believes relate to foreign intelligence, that have to do with potential terrorism investigations.

Reasons why NSA wants cell phone providers metadata

1. ID number
2. Name
3. IP address
4. MAC address
5. Location
6. Email

Ten real-valued features are computed for each observation:

* Internet Service provider
* Cell phone provider
* Bluetooth
* Home computer
* Cookie from browser
* Security camera IP
* IP Phones
* Email

**Defining a problem**

With this case study, the report also noted a 27% increase in the number of foreigners whose communications were targeted by the NSA during the year. In total, an estimated 164,770 foreign individuals or groups were targeted with search terms used by the NSA to monitor their communications, up from 129,080 on the year prior. The problem is that the accurate number of humans in the USA with a Phone or Computer system. This can affect the number of real data projected in the system or duplication.

**Defining the Target Variable**

For this case study, the target variable is very obvious, the human’s variable. This variable has two values Computer/Cell phone and Internet. Those who have internet and phone are more likely to be tracked. The report also said the NSA collected at most 434.2 million phone records on Americans, down from 534.3 million records on the year earlier. The government said the figures likely had duplicates.

**Data Understanding:**

Data shown that some 9,637 warrantless search queries of the contents of Americans’ calls, text messages, emails and other communications were conducted by the NSA during 2018, up from 7,512 searches on the year prior. Data will need to be clean and push to a train set and then it can be model.

**Data Preparation**

Based on the report, the NSA dataset needs a lot of data preparation. We can remove “Name” as it is only for identification purposes. Before removing the identity property, we can check for duplicate records and get rid of them. We can check for the data type of each column value. If needed, we can convert them to the numeric type. We can transform the Location column to binary values using the radius. We will replace the value B with 0 and value M with 1. We will inspect the distribution for all columns and will remove records with outliers if there are any. We can replace all missing values with Iterative Imputers. The Iterative Imputer will observe all other column values and compare them with the observation with the missing value. Imputer then decides the appropriate value to replace for the absent value. We can use the standard scaler to scale values as values are having different ranges. The Standard Scaling will reduce the skew impact on algorithms like the logistic regressions.

**Modeling**

Before modeling, we will split our dataset into training and test data sets. We can use an 80:20 ratio for the split as this dataset is small. We can build several prediction models for this problem. We can start with logistic regression. Then we can train other models like decision tree classifier, Random Forest classifier, K-Neighbor’s classifier, etc. While preparing the model, we can use k-fold cross-validation as the dataset is small. For the cross-validation and remove bias due to two types of observation.

**Model interpretation**

We will compare all trained models for accuracy as well as the confusion matrix by predicting values for the test dataset. As previously mentioned, the model with the small Type II error we need to select. We can finalize a model with high accuracy and small Type II error as our final model for deployment. The comparison of the ROC curve can help us deciding the best model selection.

**Model Deployment**

For all data preparation, we will use the sklearn. I will be applying the same operations on unseen data before applying model prediction. Because this model is used is from a telecommunication system like ATT, Verizon and Cox internet, we have to observe the model results frequently. We also need to retrain the model regularly as we are starting with a small dataset. It will also help us in accommodating the accurate of data pulled in.

**References**

<https://www.grin.com/document/308419-> Information security for national security: The Snowden and NSA case study

<https://techcrunch.com/2019/04/30/nsa-surveillance-spike/-> NSA says warrantless searches of Americans’ data rose in 2018

https://www.kaggle.com/cityofLA/los-angeles-311-call-center-tracking-data- Los Angeles 311 Call Center Tracking Data

[https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html sklearn.linear\_model.LogisticRegression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html%20sklearn.linear_model.LogisticRegression)

https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.train\_test\_split.html- sklearn.model\_selection.train\_test\_split