# Opioid Detection in US Mail Stream

By: Danial Mirza & Rihab Rebai

Faculty Advisor: Negin Golrezaei

## PROBLEM STATEMENT

***1M packages*** enter the US mail stream from abroad everyday. Only a small number of these will contain opioids. With a budget to open and inspect only 30 packages daily, ***which should we open?***

### Goals

1. Develop a robust classification model which detects inbound packages containing opioids with high confidence among the riskiest packages in order to flag them for inspection.
2. Construct a large-scale graph to examine relationships between past seizures as a proof-of-concept, launching internal capabilities for graph analysis at USPS.

### Metric of Choice: 1% Lift Rate

We aim to maximize the 1% lift rate of the model defined as follows:

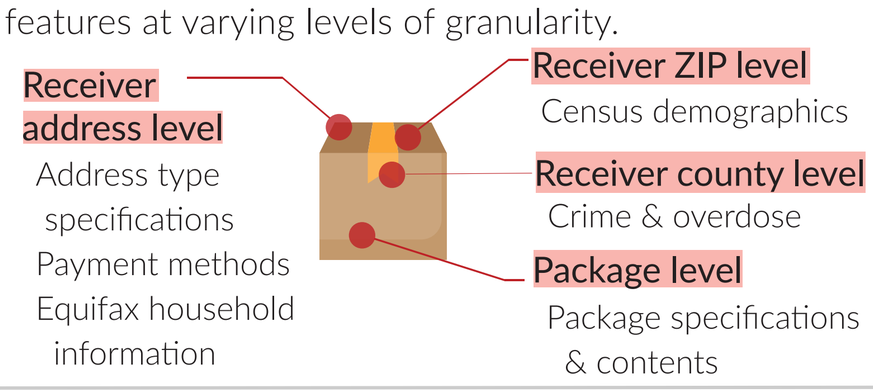
$$\frac{\#\text{ true positives in top 1\% of suspect packages}}{\text{total \# of packages in top 1\%}} \times \frac{}{\%\text{ positives in total packages}}$$

This metric quantifies how well we are predicting true positives among the most risky packages (i.e. the ones we intend to open and inspect). It is normalized against the proportion of opioids in the dataset.

## DATASET

**Seizures dataset**

120k packages deemed suspect and opened between 2015 - 2019. Contains package contents and ~330 features at varying levels of granularity.



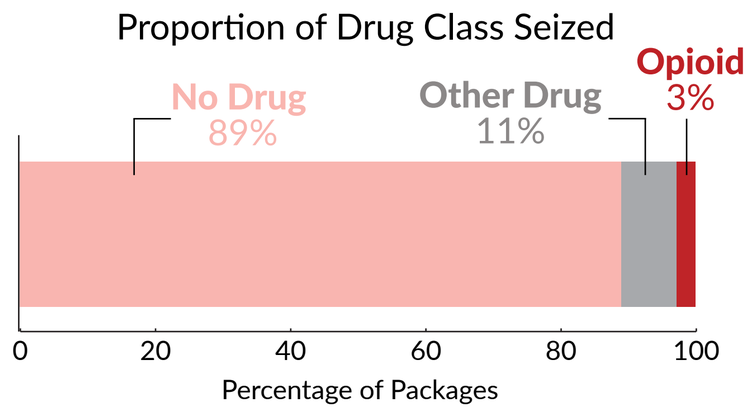

**Additional datasets**

**Mail Forwarding**
Mail forwarded from one address to another. 130M requests over 2016-2020.

**Incoming Packages**
International inbound packages, similar features to seizures. The true contents are unknown. ~1M packages daily.

**Key Identifiers:** label number and address

*Opioids are only 3% of seized packages*




## CLASSIFICATION MODEL

*We train and evaluate our model using the seizures data*

**We propose two modelling outcomes:**

1) Opioid vs. Regular Mail
2) Other Drug vs. Regular Mail

In order to better understand which features uniquely identify opioids.

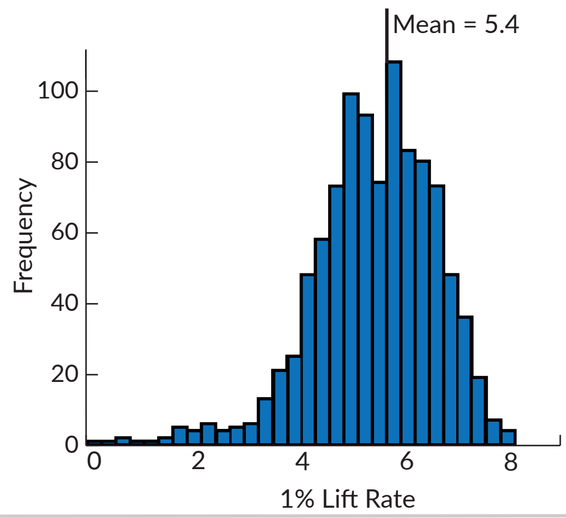
**Needle in a Haystack**
Due to the severe class imbalance, we experiment with sampling techniques to improve performance.

**Missing Package Data**
Sometimes, package specifications e.g. weight are not available - we 'fill in the blanks' using imputation.

**Highly Correlated Features**
Many of the census variables are duplicated or correlated. We apply PCA to reduce noise in the dataset.

**Generalization to New Addresses**
The observed seizures are sparse with the vast majority of addresses unobserved. Our model is at risk of repeatedly flagging the same locations. We split our data such that the same ZIP code does not appear in both train and test to evaluate performance of the model on unseen addresses.

*Our best model, a random forest, achieves an average 1% lift rate of 5.4*




## GRAPH ANALYTICS

**Step 1** Starting with mail forwarding data, we develop an algorithm which finds chains of movement in order to track mail-pieces.

| Old Address | New Address | Move Date | Name | Unique Name |
|---|---|---|---|---|
| 1111 | 2222 | Jun-1-2016 | John Doe | johndoe1 |
| 2222 | 3333 | Jul-1-2017 | John Doe | |
| 2222 | 6666 | Feb-4-2019 | John Doe | johndoe2 |
| 4444 | 5555 | Mar-9-2020 | John Doe | johndoe3 |

**Step 2** We develop an algorithm to identify associations where there is a shared address. We find 1BN associations.

**Step 3** Using names as nodes and associations as edges, we construct a graph. This graph can be used to find individuals who know each other through first, second, and third degree connections or clusters.

### Results & Deliverables

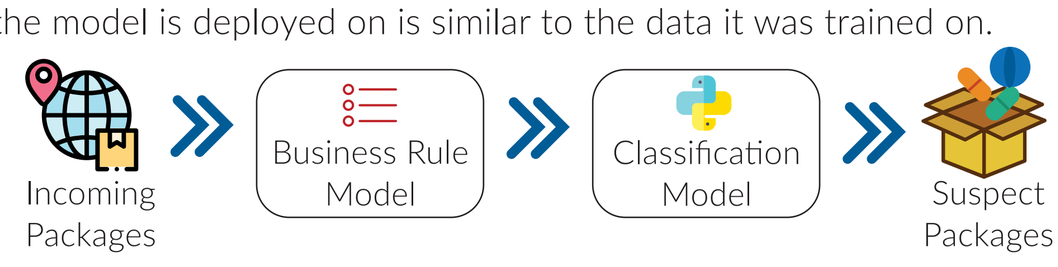
- ☆ 3 novel use cases for graph analytics at USPS
- ☆ Fully scaled graph for known associates use case in Spark
- ☆ One hour graph analytics seminar to 25 members of Advanced Analytics group, inspiring new project ideas

## DEPLOYING MODEL

A final consideration for deploying the model is translating it from the seizures dataset (on which it was trained and evaluated) to the incoming packages data.

The seizures are a biased sample from historical incoming packages, selected using a 'business rules' model.

Thus, we propose the following approach to ensure that the data the model is deployed on is similar to the data it was trained on.




## BUSINESS IMPACT

Our model achieves an **out-of-sample lift rate of 5.4,** promising to perform over 5x better than the current business rules model if deployed on top of it, thus intercepting more opioids in the mail.

**Unifying business rules and classification model**, traditionally seen as competing models.

**Inspiring new use-cases for graph analysis** and providing a strong starting point for any future work through detailed documentation.

**Quickening response to changing criminal behavior** with an end-to-end pipeline enabling model to be retrained as new seizures data becomes available.