William Rubin
DATA606, Spring 2020

PART 3 SUMMARY: Predictive Modeling of Opioid Prescription Fraud by Centers for Medicare and Medicaid Services (CMS) Providers

The main goal of this part was to prepare the data for use with machine learning algorithms and predict Medicare Part D Providers which may be fraudulently prescribing opioids. In order to accomplish this objective, a labeled dataset was constructed which contained all CMS Providers identified in the Part D Prescriber Summary Table as having provided opioids to patients. The opioid provider dataset was then augmented to include a classifier to indicate fraudulent / legitimate practices in addition to a indication of level of certainty.

The following tasks were completed to create the CMS Opioid Provider Labeled Dataset (depicted in Figure 1 below), which will be utilized in Part 4 to predict the likelihood of prescription related fraud:

- Developed Data Pipeline for Fraudulent Provider Identification
    - *Data Collection*: DOJ Press Release website, filtered for opioid-related cases, was scraped utilizing Scrapy to compile a corpus of text associated with CMS providers formally charged, plead guilty to, or convicted of fraudulent activity. When executed on 2020-03-11, the data collection Python code retrieved **120 press releases** from the DOJ website.
    - *Entity / Location Extraction*: Dataset of each scraped opioid-related DOJ Press Release text was cleansed and spaCy (NLP Python Library) was utilized to extract entities from the text. The spaCy library supports extraction of numerous entities, for my purposes the entities extracted included: ORG (organization), PERSON (names of people), GPE (country, city, state), and LOC (non-GPE location).
    - *NPI Resolution*: For all people (first and last name) and states extracted from each scraped press release, API calls were constructed to lookup and retrieve any officially assigned identifiers from the NPPES NPI Registry matching the parameters. Of the **overall 856** matches found, **260 resulted in exact matches** and were then included in the initial known fraudulent provider dataset. For Part 4 of this project, this process will be improved to handle instances when multiple potential matches are returned from the NPPES NPI Registry API Lookup.
- Performed Final EDA of Part D Prescriber Summary Table Data
    - *Opioid Provider Dataset*: Created aggregated dataset of **704,463 providers** based on NPI which reported values for opioid or long-acting opioid characteristics. Opioid provider analysis conducted was based upon records grouped by NPI across the five years contained within the Summary Table Data. This decision was made due to time constraints of extracting relevant timeframes in which the fraud occurred from DOJ Press Releases using NLP.
    - *Likely Legitimate Provider Determination*: Analyzed opioid provider dataset to identify averages and outliers in terms of opioid and long-acting opioid characteristics (Beneficiary Count, Claim Count, Prescription Rate, Drug Cost,

and Day Supply). Of note, the assumptions necessary to determine likely legitimate providers was not initially taken into consideration but is necessary to properly classify providers in the labeled dataset. This determination will be revised prior to execution of predictive models in Part 4.

- Predictive Modeling Preparation
  - *Establish Classifiers*: Determined the following two classifiers to be used in the creation of the labeled dataset: "Provider Classification" will be used to distinguish between "Legitimate" (value = 0) and "Fraudulent" (value = 1); "Certainty Level" will be used to identify "Unknown" (value = 0) and "Known" (value = 1). Based on these classifiers, records with values of (1,1) would indicate fraudulent providers identified though the data mining process and those with values of (0,0) would indicate providers determined to be likely legitimate.
  - *CMS Part D Opioid Provider Labeled Dataset Creation*: Combined Opioid Provider Dataset with classified listing of fraudulent and likely legitimate providers determined in prior steps to compile labeled data. Classifiers for remaining providers will be left blank to indicate "unknown" status.
  - *Model Construction*: Based upon initial research and evaluation of the scikit-learn algorithm cheat sheet, focus will be placed on implementation of Linear Support Vector Classification (SVC) and Naive Bayes. The analysis will be performed using Orange for rank and feature selection, split of data between testing and training, and algorithm execution. Other algorithms will also be explored and evaluated in terms of applicability and accuracy to produce the most statistically significant results.

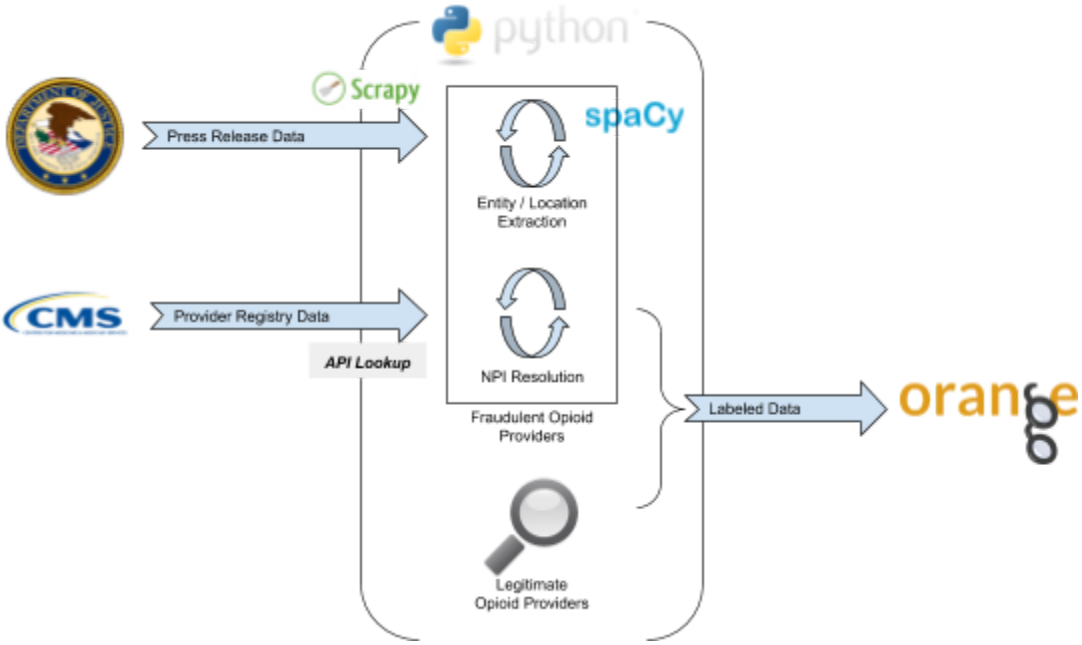

Figure 1. CMS Part D Opioid Provider Labeled Dataset Process

The following table provides details pertaining to the additional libraries and products utilized in this project:

Table 1. CMS Part D Opioid Provider Labeled Dataset Technology Stack

Product / Library Name	**URL**	**Purpose**
Scrapy	https://scrapy.org	Open-source Python library which provides a framework to facilitate the extraction of press release text data from the DOJ Website.
spaCy	https://spacy.io/	Open-souce Python library which provides Natural Language Processing (NLP) capabilities and utilized to perform entity extraction (names and locations) from corpus of DOJ Press Release scraped text.
Orange	https://orange.biolab.si/	Open source Python based environment providing robust machine learning and data visualization capabilities for executing predictive analysis.