Open Payments Analysis

Sergio Mastrogiovanni

October 22, 2019

About me

... a Data Scientist AI evangelist and data storyteller

... a Professor of Intelligent Automation at New York University

... a Consultant with 20+ years in Continuous Improvement exp.

... an Innovation Coach entrepreneur and global leader.

... an Analytics Researcher at NYU Center for Sustainable Business

Outline

- Background
- Approach
- Python Model
- Analysis
- Next Steps

Background

Open Payments:

- Disclosure program managed by the Centers for Medicare & Medicaid Services (*CMS*).
- Promotes transparency and accountability.
- Helps consumers understand the financial relationships between pharmaceutical and medical device industries, and physicians and teaching hospitals.
- Financial relationships may include consulting fees, research grants, travel reimbursements, and payments made from the industry to medical practitioners.
- Data was taken from CMS site (2017): https://www.cms.gov/OpenPayments/Explore-the-Data/Dataset-Downloads.html

Background

Data Set:

- Annual data collection (2017). ~7Gb

- 4 Files:

    1. <u>General Payments (*OP_DTL_GNRL_PGYR2017_P06282019.csv*)</u>: Payments or other transfers of value made that are not in connection with a research agreement or research protocol.
    2. <u>Research Payments (*OP_DTL_RSRCH_PGYR2017_P06282019.csv*)</u>: Payments or other transfers of value made in connection with a research agreement or research protocol.
    3. <u>Physician Ownership or Investment Interest Information (*OP_DTL_OWNRSHP_PGYR2017_P06282019.csv*)</u>: Information about physicians who hold an ownership or investment interest in an applicable manufacturer or applicable GPO or who have an immediate family member holding such interest.
    4. <u>Removed/Deleted records (*OP_REMOVED_DELETED_PGYR2017_P06282019.csv*)</u>: Payments removed from previous analysis.

- Stakeholoders:

Patients and Consumers | **Physicians and Entities** | **Companies and GPOs*** | **Researchers**

*Group Purchasing Organisations

Approach

Tools

alteryx + python™ + tableau

Exploratory data analysis | **Modeling** | **Data Visualizations**

Approach

70% ***of time spent (~4 hours)***

Business Understanding
- Understand Objective
- CMS site and process
- Determine Data Mining Goals

Data Understanding
- Download data
- Describe Data
- Explore Data
- Verify Data Quality

Data Preparation
- Clean Data
- Construct Data
- Integrate & Format Data

Modeling
- Logistic Regression
- Generate Test Design
- Build Model
- Assess Model

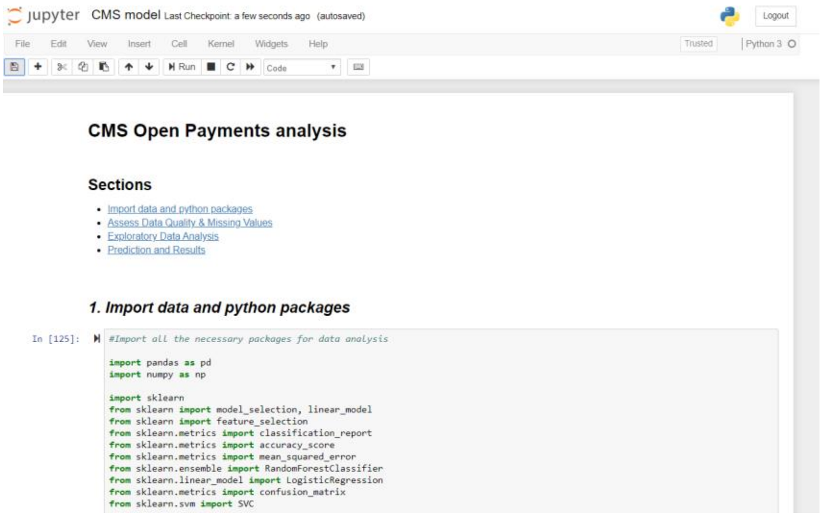
Evaluation
- Evaluate Results
- Review Process
- Determine Next Steps

Reporting
- Create visualizations
- Create PowerPoint
- Address questions
- Review Presentation

Python Model

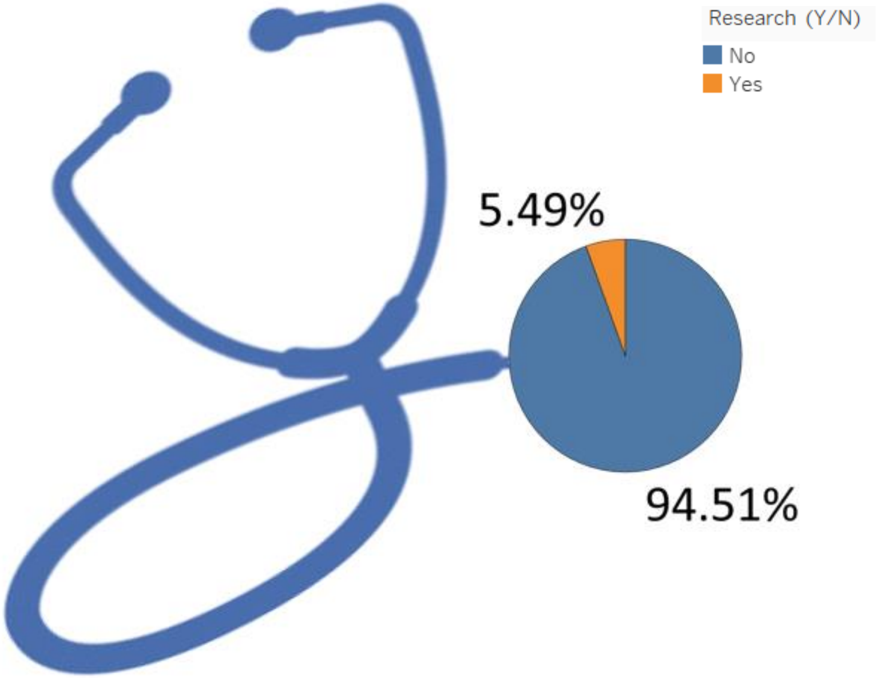
1. Import Data and Python packages
2. Assess Data Quality & Missing Values
3. Exploratory Data Analysis
4. Prediction and Results

https://github.com/sergiomastro/CMS/blob/master/CMS%20model.ipynb



Analysis

Exploratory data analysis



Research projects represent:

- 5.49% of the payments
- 653,488 payments made in 2017
- $5.10 Billion

Analysis

Exploratory data analysis

Total Dollar Value

$8.9 Billion

Total Records Published

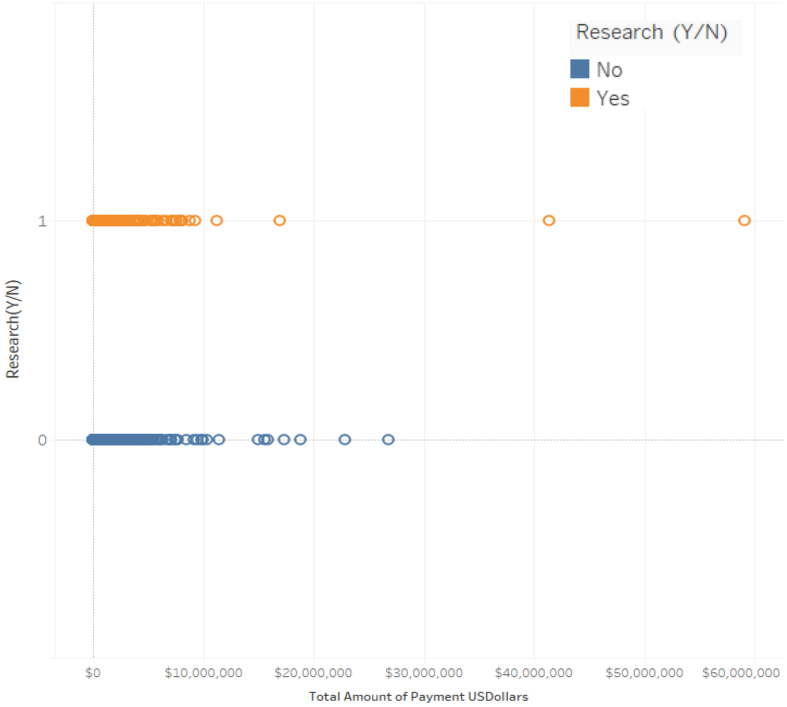
11.89 Million*

$ General Payments	Amount **$2.90 Billion**	Payments* **11.24 Million**
Research Payments	Amount **$5.10 Billion**	Payments* **653,488**
Value of Ownership	Amount **$976.93 Million**	Payments* **2,840**

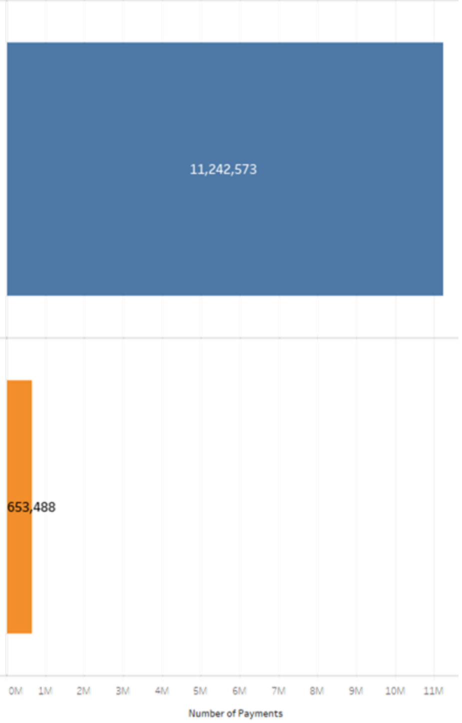
*Each Payment has a unique **Record ID**

Analysis

Exploratory data analysis



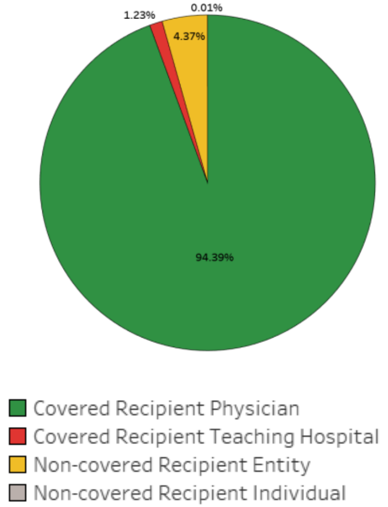
Total Amount of Payment USDolla..	Avg. Total Amount of Payment USDollars	Min. Total Amount of Payment USDollars	Max. Total Amount of Payment USDollars	Std. dev. of Total Amount of Payment USDollars
$2,900,747,465	$258	$0	$26,842,366	$21,586
$5,097,044,013	$7,800	$0	$59,104,128	$111,479

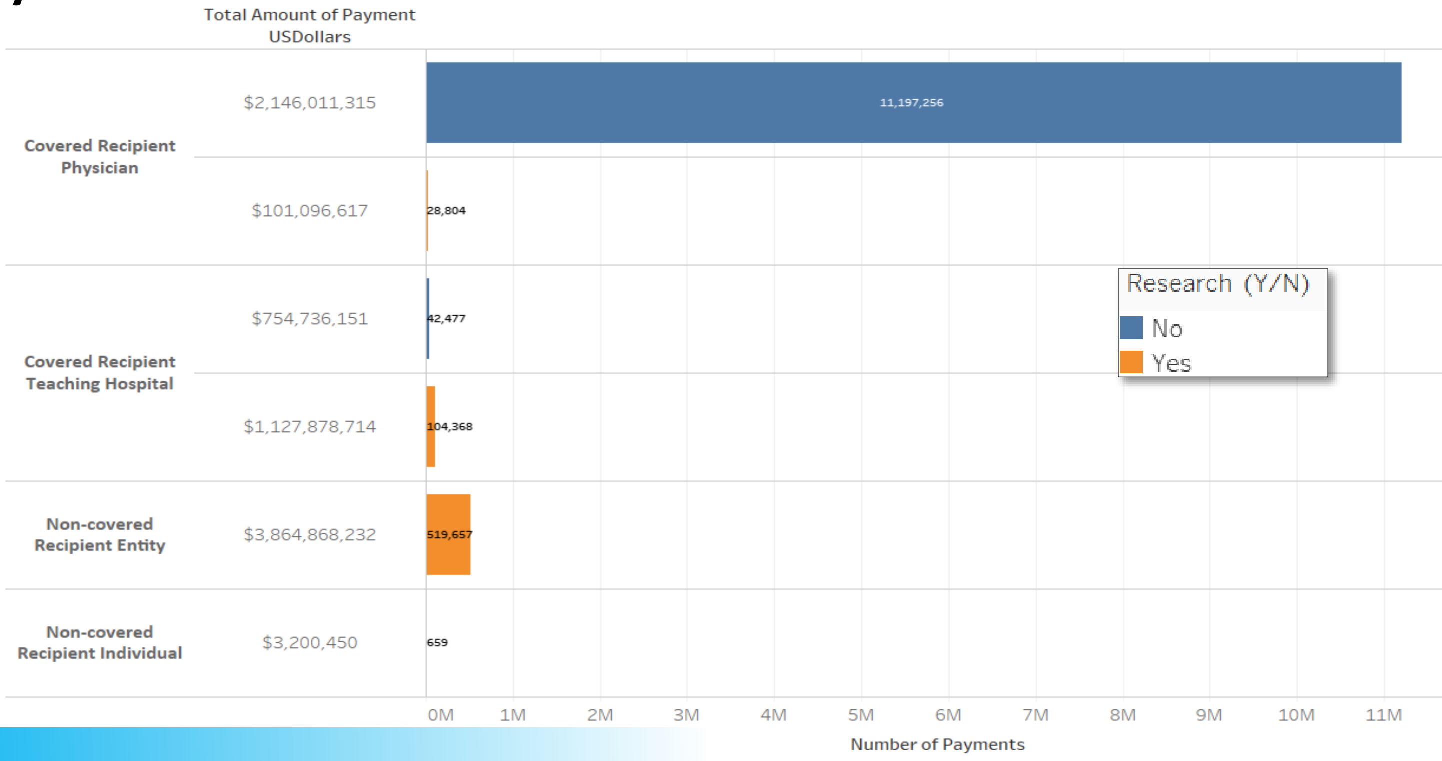


Analysis

Exploratory data analysis

Type of payments:

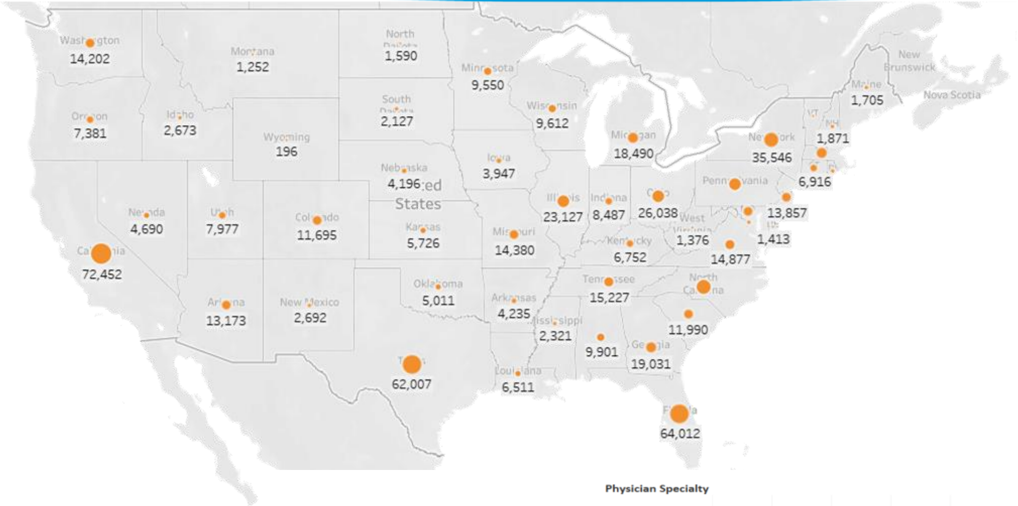




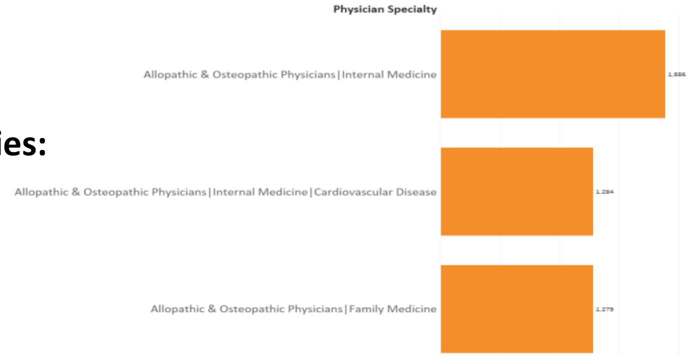
Analysis

Research Projects:

Recipient Country	Number of Records	Total Amount of Payment USDollars	Avg. Total Amount of Payment USDollars	Min. Total Amount of Payment USDollars	Max. Total Amount of Payment USDollars	Std. dev. of Total Amount of Payment USDollars
United States	652,268	$5,090,301,970	$7,804	$0	$59,104,128	$111,579
Null	659	$3,200,450	$4,857	$15	$337,500	$15,238
Canada	96	$3,026,142	$31,522	$20	$195,750	$54,826
Great Britain (Uk)	441	$249,236	$565	$10	$30,483	$2,314
Belgium	16	$245,809	$15,363	$6	$45,003	$13,052
Germany	1	$6,909	$6,909	$6,909	$6,909	
United States Minor Outlying Islands	3	$5,437	$1,812	$279	$4,020	$1,959
Australia	1	$4,336	$4,336	$4,336	$4,336	
Poland	1	$3,113	$3,113	$3,113	$3,113	
Denmark	1	$563	$563	$563	$563	
Japan	1	$48	$48	$48	$48	



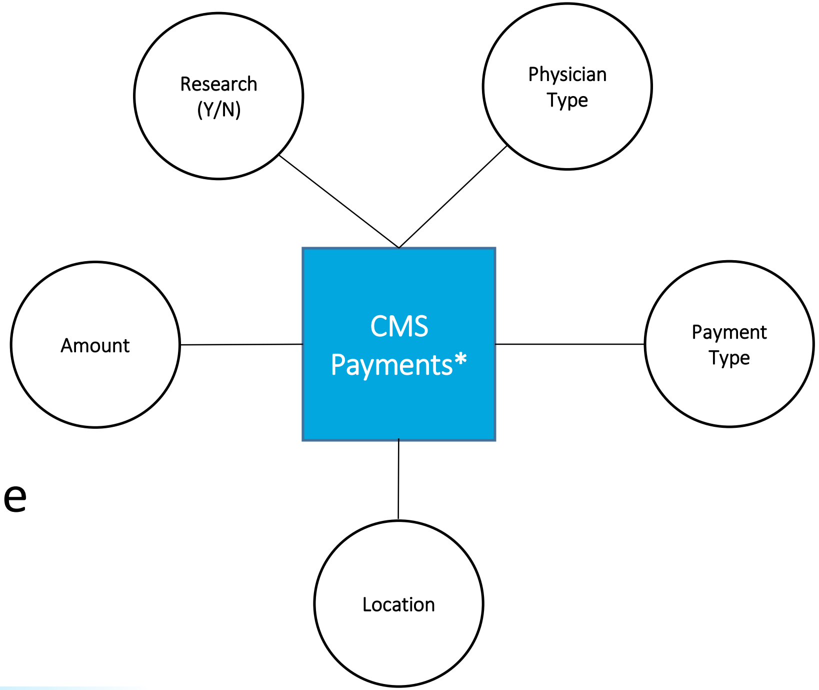
Top 3 Specialties:



Analysis

Data Prep & Sampling:

- Remove Deleted & Ownership records
- Remove nulls and correlated records
- Remove extra spaces
- Remove useless, redundant and noisy features (columns)
- Label *Research* records & join datasets
- Sampling:
  - Select a statistically significant sample
  - Confidence level: 95%, CI: .1
  - Sample size: ***883,282***



*Each Payment has a unique **Record ID**

Analysis

Model:

- Unsupervised learning
  - Logistic Regression
  - Decision Tree
  - Random Forest
- Used observation to sample the relevant features from the dataset.
- Imbalanced dataset (number of observations for research payments are significant less than others).
- Didn’t use Receiver Operating Characteristic(ROC) to find out the true positive rate over the false positive rate.

Next Steps

- Establish a baseline to which compare results later.
- Explore the frequent of the payments (weekly).
- Include regularization (lasso regularization) to reduce weights for features that are not significant to zero.
- Use a fully connected neural network: we have an important amount of data that would make Neural Networks work best.
- Lastly, do PCA to reduce the dimensionality of the data.

Lessons for future data collection

- Find a better technique for data sampling.
- Use the experimental study rather than observation study.
- Collection of unbiased datasets with a lesser class imbalances.
- Look for the causation rather than correlation of independent and dependent variables.

Data set attributes:

- Analyse common features across files.
- Review the data dictionary to understand the relationships.
- Remove correlated fields (e.g. address/city, name/last name, etc.).
- ID and name variables that must be discarded to ensure best accuracy and efficient computation of the algorithms used.
- Distinguish factors for payments:
    - Location
    - Type of Payment
    - Type of Physician
    - Total amount USD

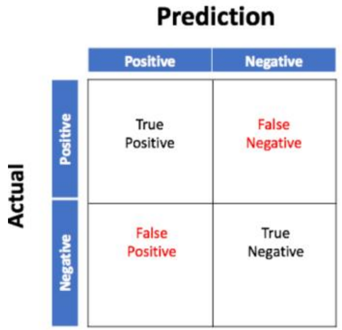
Pitfalls

- Class imbalance: accuracy considered to be the best matrix to evaluate the results performance of the algorithm.

An algorithm which always predicts 0 (payment not for research purpose), the model would still give 99% accuracy because the 99% of the data is the one having the class 0.

We know that this is the worst algorithm which always predicts 0 no matter what, therefore we use and rely on Precision/Recall rather than ROC or Accuracy. Therefore, the f1-score, which is the arithmetic mean of precision and recall will cater this situation and give the correct output.

- Assumptions: Multicollinearity, Heteroscedasticity, normality.
- Outliers and overfitting.



Thank you