Medicare Fraud

GREYSON GARDNER
JASON HARRIS
CLERISSE LEMKE
JIANI XU
JAYE WALL

GEORGETOWN UNIVERSITY

COHORT 18, SPRING 2020

CAPSTONE PROJECT

CERTIFICATE IN DATA SCIENCE

Background

Business Questions and Hypothesis

Initial

- Can We Predict Fraudulent Activity Among Healthcare Providers Based On Historical Exclusions From The List Of Excluded Individuals And Entities (LEIE)?

Reframed

- Can We Predict Providers That Will Be Excluded Based On The Historical Data From The List Of Excluded Individuals And Entities (LEIE)?

Why are Health Providers on List of Excluded Individuals and Entities (LEIE)

Authorities: Pursuant to section 1128 of the Social Security Act (Act) (and from Medicare and State health care programs under section 1156 of the Act)

Exclusions are imposed for a number of reasons:

- **Mandatory exclusions:**
  - Participation in all Federal health care programs individuals and entities convicted of the following types of criminal offenses
    - Patient abuse or neglect; felony convictions for other health care-related fraud, theft, or other financial misconduct; and felony convictions relating to unlawful manufacture, distribution, prescription, or dispensing of controlled substances
- **Permissive exclusions:**
  - Individuals and entities on a number of grounds, including (but not limited to) misdemeanor convictions related to health care fraud other than Medicare or a State health program, fraud in a program (other than a health care program) funded by any Federal, State or local government agency;
  - Misdemeanor convictions relating to the unlawful manufacture, distribution, prescription, or dispensing of controlled substances; suspension, revocation, or surrender of a license to provide healthcare for reasons bearing on professional competence, professional performance, or financial integrity; provision of unnecessary or substandard services; submission of false or fraudulent claims to a Federal health care program; engaging in unlawful kickback arrangements; defaulting on health education loan or scholarship obligations; and controlling a sanctioned entity as an owner, officer, or managing employee.

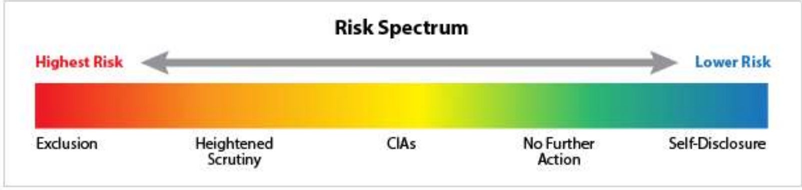
Risk Categories

Highest Risk - Exclusion

Parties that OIG determines present the highest risk of fraud will be excluded from Federal healthcare programs to protect those programs and their beneficiaries. Excluded individuals and entities are listed in OIG's Exclusions Database.

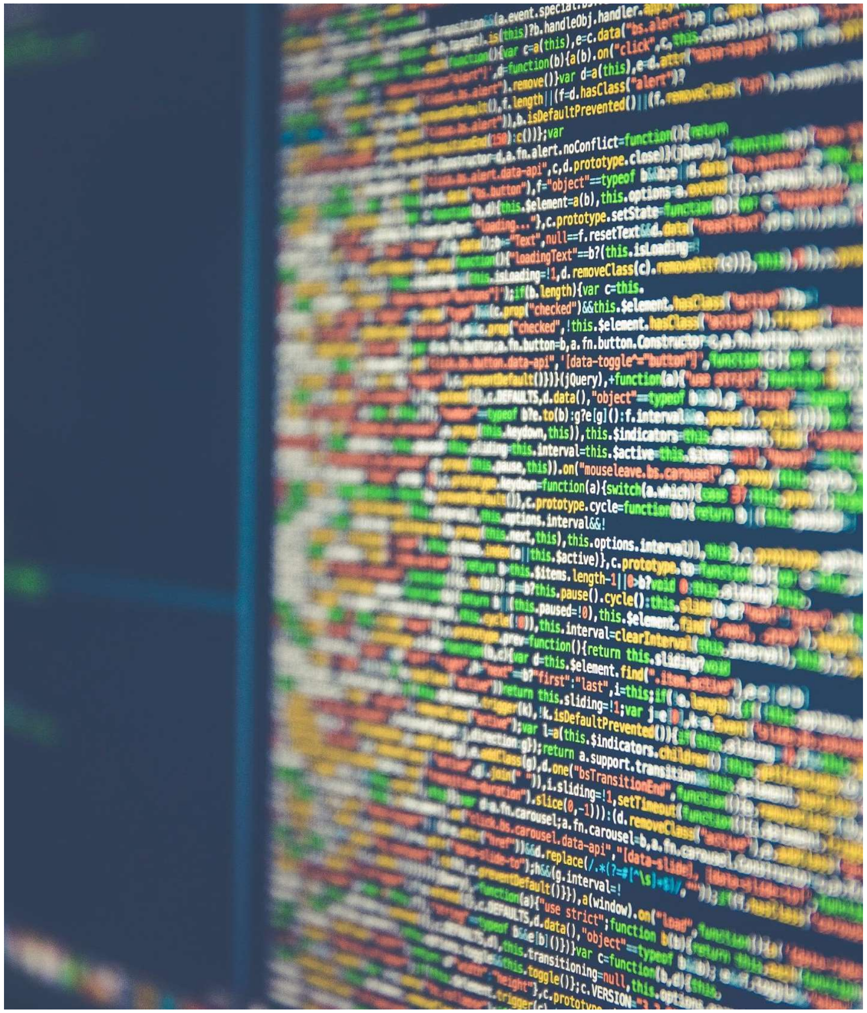
Fraud Risk Indicator

OIG assessment of future risk posed by persons who have allegedly engaged in civil healthcare fraud.



What's at RISK?
List of Excluded Individuals and Entities (LEIE)

Data



THE DATA SETS

- List of Excluded Individuals and Entities (LEIE)
  - All current exclusions data
- Medicare Provider Utilization and Payment Data
  - Part D Prescriber Public Use File (PUF)
  - CY 2017 Prescriber Summary Table

THE DATA SETS JOINS, FEATURES AND OUTCOMES

Joins

- National Provider Identifier(NPI)Number
- Name and other identification fields

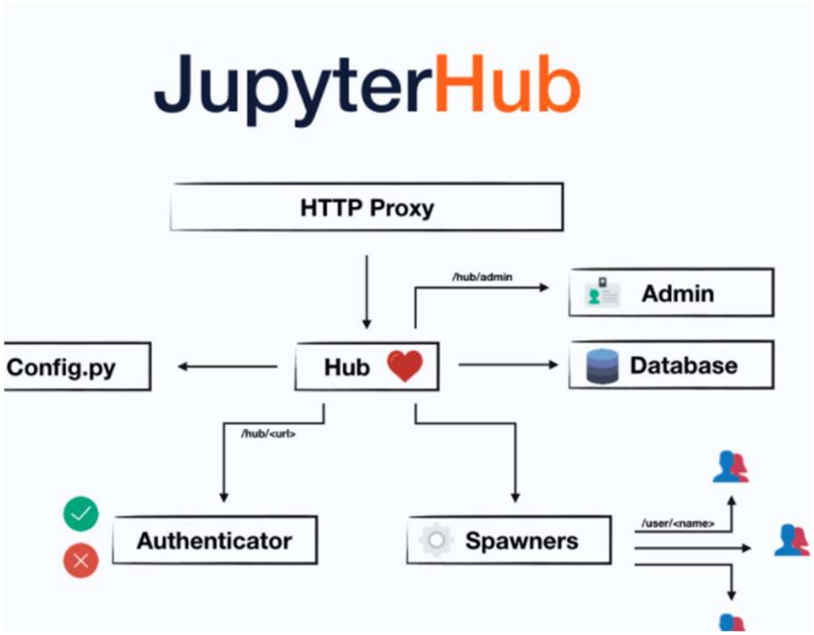
Features

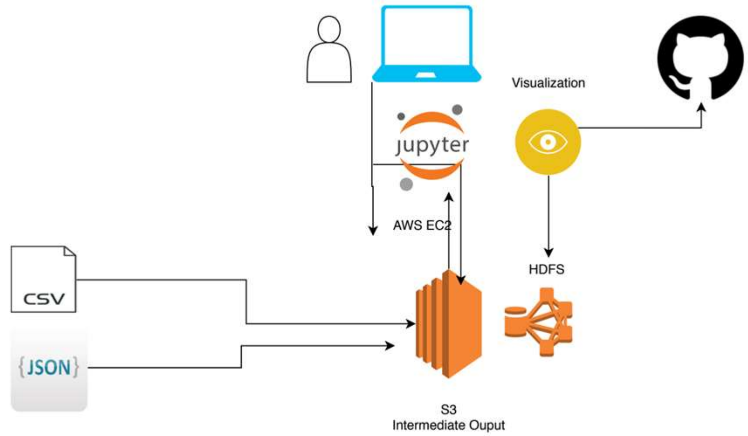
- Drug costs
- Drug types
- Number of claims

Outcome

- Medicare Fraud
- Medicare Exclusions

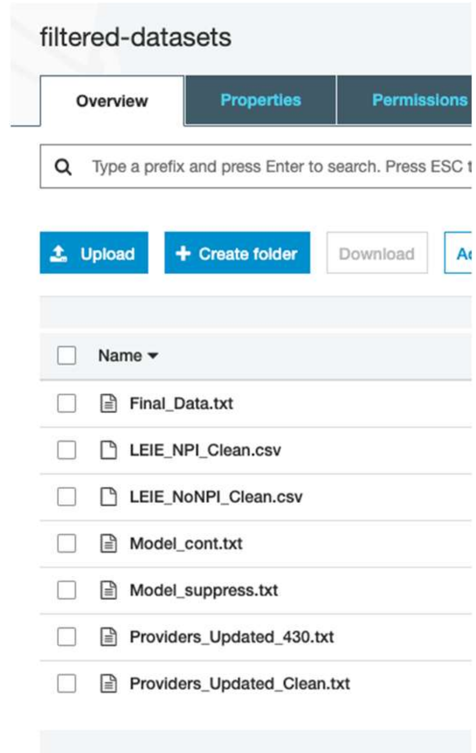
Data Ingestion Process





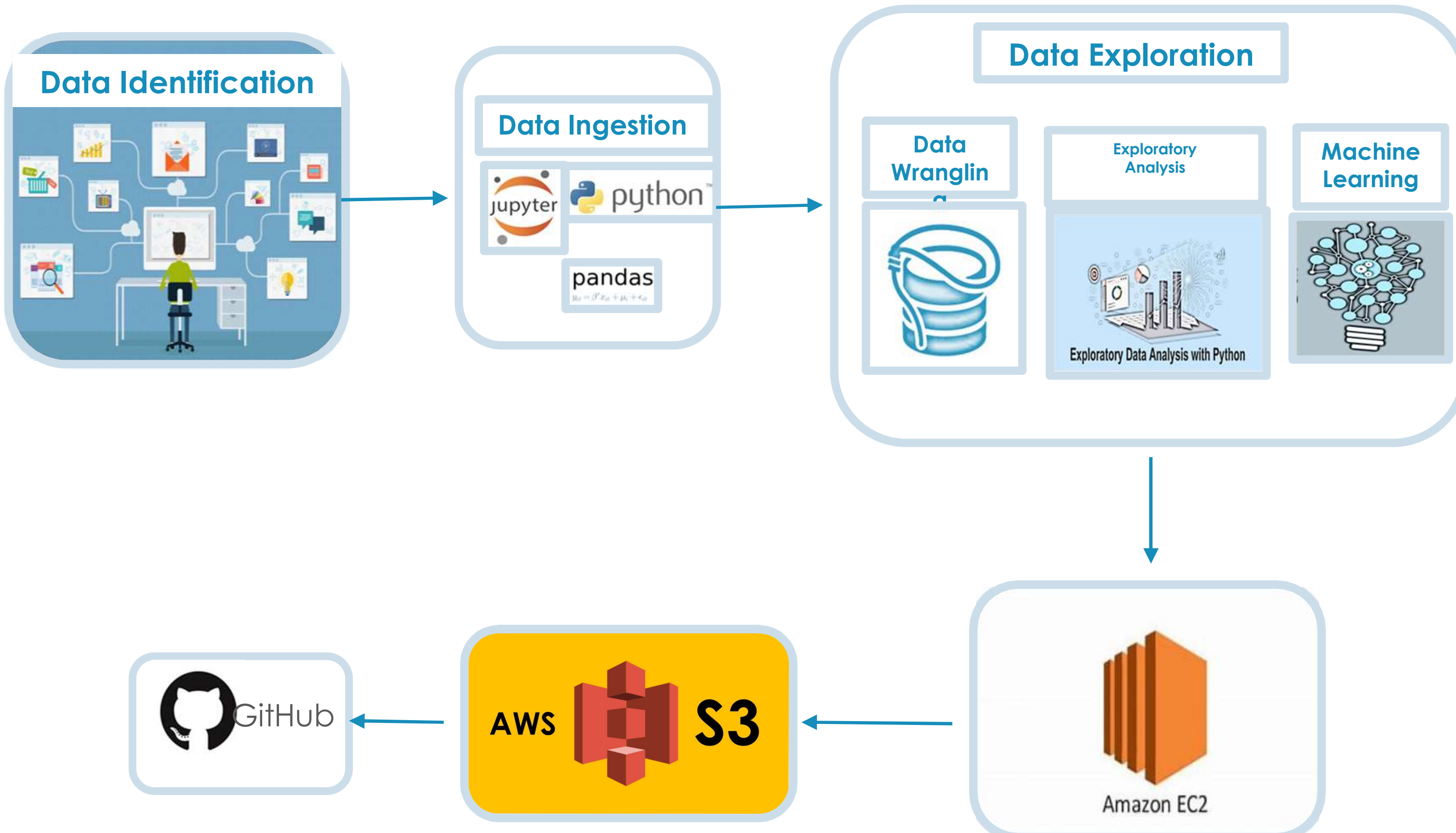
- Raw data will be loaded into s3 buckets
- Raw data inputs & Intermediate output will be stored in s3
- JupyterHub will be used by group for wrangling, ingestion, eda.
- Once Tables are normalized, they will be placed into RDS
- From this point they will be used for visualization

Architecture



Simple Storage Solution (S3)

Data Ingestion Workflow Process



DATA WRANGLING

Snapshot of The Data

```
Index(['npi', 'nppes_provider_last_org_name', 'nppes_provider_first_name',
       'nppes_provider_mi', 'nppes_credentials', 'nppes_provider_gender',
       'nppes_entity_code', 'nppes_provider_street1', 'nppes_provider_street2',
       'nppes_provider_city', 'nppes_provider_zip5', 'nppes_provider_zip4',
       'nppes_provider_state', 'nppes_provider_country',
       'specialty_description', 'description_flag',
       'medicare_prvdr_enroll_status', 'total_claim_count',
       'total_30_day_fill_count', 'total_drug_cost', 'total_day_supply',
       'bene_count', 'ge65_suppress_flag', 'total_claim_count_ge65',
       'total_30_day_fill_count_ge65', 'total_drug_cost_ge65',
       'total_day_supply_ge65', 'bene_count_ge65_suppress_flag',
       'bene_count_ge65', 'brand_suppress_flag', 'brand_claim_count',
       'brand_drug_cost', 'generic_suppress_flag', 'generic_claim_count',
       'generic_drug_cost', 'other_suppress_flag', 'other_claim_count',
       'other_drug_cost', 'mapd_suppress_flag', 'mapd_claim_count',
       'mapd_drug_cost', 'pdp_suppress_flag', 'pdp_claim_count',
       'pdp_drug_cost', 'lis_suppress_flag', 'lis_claim_count',
       'lis_drug_cost', 'nonlis_suppress_flag', 'nonlis_claim_count',
       'nonlis_drug_cost', 'opioid_claim_count', 'opioid_drug_cost',
       'opioid_day_supply', 'opioid_bene_count', 'opioid_prescriber_rate',
       'la_opioid_claim_count', 'la_opioid_drug_cost', 'la_opioid_day_supply',
       'la_opioid_bene_count', 'la_opioid_prescriber_rate',
       'antibiotic_claim_count', 'antibiotic_drug_cost',
       'antibiotic_bene_count', 'antipsych_ge65_suppress_flag',
       'antipsych_claim_count_ge65', 'antipsych_drug_cost_ge65',
       'antipsych_bene_ge65_suppress_flg', 'antipsych_bene_count_ge65',
       'average_age_of_beneficiaries', 'beneficiary_age_less_65_count',
       'beneficiary_age_65_74_count', 'beneficiary_age_75_84_count',
       'beneficiary_age_greater_84_count', 'beneficiary_female_count',
       'beneficiary_male_count', 'beneficiary_race_white_count',
       'beneficiary_race_black_count', 'beneficiary_race_asian_pi_count',
       'beneficiary_race_hispanic_count', 'beneficiary_race_nat_ind_count',
       'beneficiary_race_other_count', 'beneficiary_nondual_count',
       'beneficiary_dual_count', 'beneficiary_average_risk_score'],
      dtype='object')
```

Medicare Provider Utilization and Payment Data (Part D Prescriber):

- 84 columns and 1,162,898 rows
- Numeric and object data types
- Null values

```
provider.head()
```

	npi	nppes_provider_last_org_name	nppes_provider_first_name	nppes_provider_mi	nppes_credentials	n
0	1003000126	ENKESHAFI	ARDALAN	NaN	M.D.	M
1	1003000142	KHALIL	RASHID	NaN	M.D.	M
2	1003000167	ESCOBAR	JULIO	E	DDS	M
3	1003000175	REYES-VASQUEZ	BELINDA	NaN	D.D.S.	F
4	1003000282	BLAKEMORE	ROSIE	K	FNP	F

```
df.columns
```

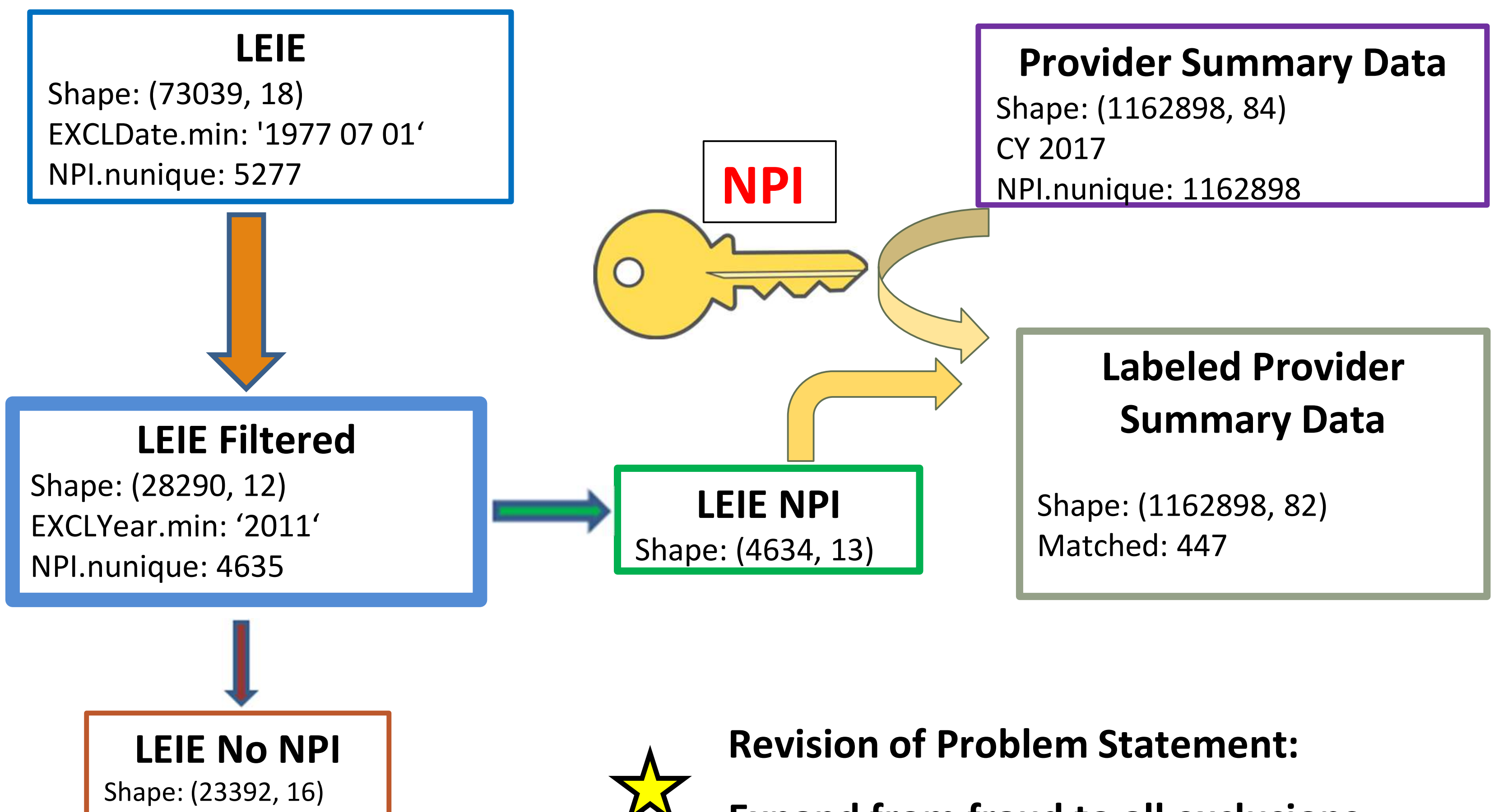
```
Index(['LASTNAME', 'FIRSTNAME', 'MIDNAME', 'BUSNAME', 'GENERAL', 'SPECIALTY',
       'UPIN', 'NPI', 'DOB', 'ADDRESS', 'CITY', 'STATE', 'ZIP', 'EXCLTYPE',
       'EXCLDATE', 'REINDATE', 'WAIVERDATE', 'WVRSTATE'],
      dtype='object')
```

```
df.head(10)
```

	LASTNAME	FIRSTNAME	MIDNAME	BUSNAME	GENERAL	SPECIALTY	UPIN	NPI	DOB	ADDRESS	CITY	STATE	ZIP
0	NaN	NaN		14 LAWRENCE AVE PHARMACY	PHARMACY	NaN	NaN	0	NaN	14 LAWRENCE AVENUE	SMITHTOWN	NY	11787
1	NaN	NaN		143 MEDICAL EQUIPMENT CO	DME COMPANY	DME - OXYGEN	NaN	0	NaN	701 NW 36 AVENUE	MIAMI	FL	33125
2	NaN	NaN		184TH STREET PHARMACY CORP	OTHER BUSINESS	PHARMACY	NaN	1922348218	NaN	69 E 184TH ST	BRONX	NY	10468
3	NaN	NaN		1951 FLATBUSH AVENUE PHARMACY	PHARMACY	NaN	NaN	0	NaN	1951 FLATBUSH AVE	BROOKLYN	NY	11234
4	NaN	NaN		1ST COMMUNITY HEALTH CTR, LTD	CLINIC	NaN	NaN	0	NaN	3138 W CERMAK ROAD	CHICAGO	IL	60623
5	NaN	NaN		1ST REHABILITATION OF PORT ST	MANAGEMENT SVCS CO	NaN	NaN	0	NaN	C/O 3659 MAGUIRE BLVD	ORLANDO	FL	32803

Snapshot of the data
List of Excluded Individuals/Entities (LEIE)

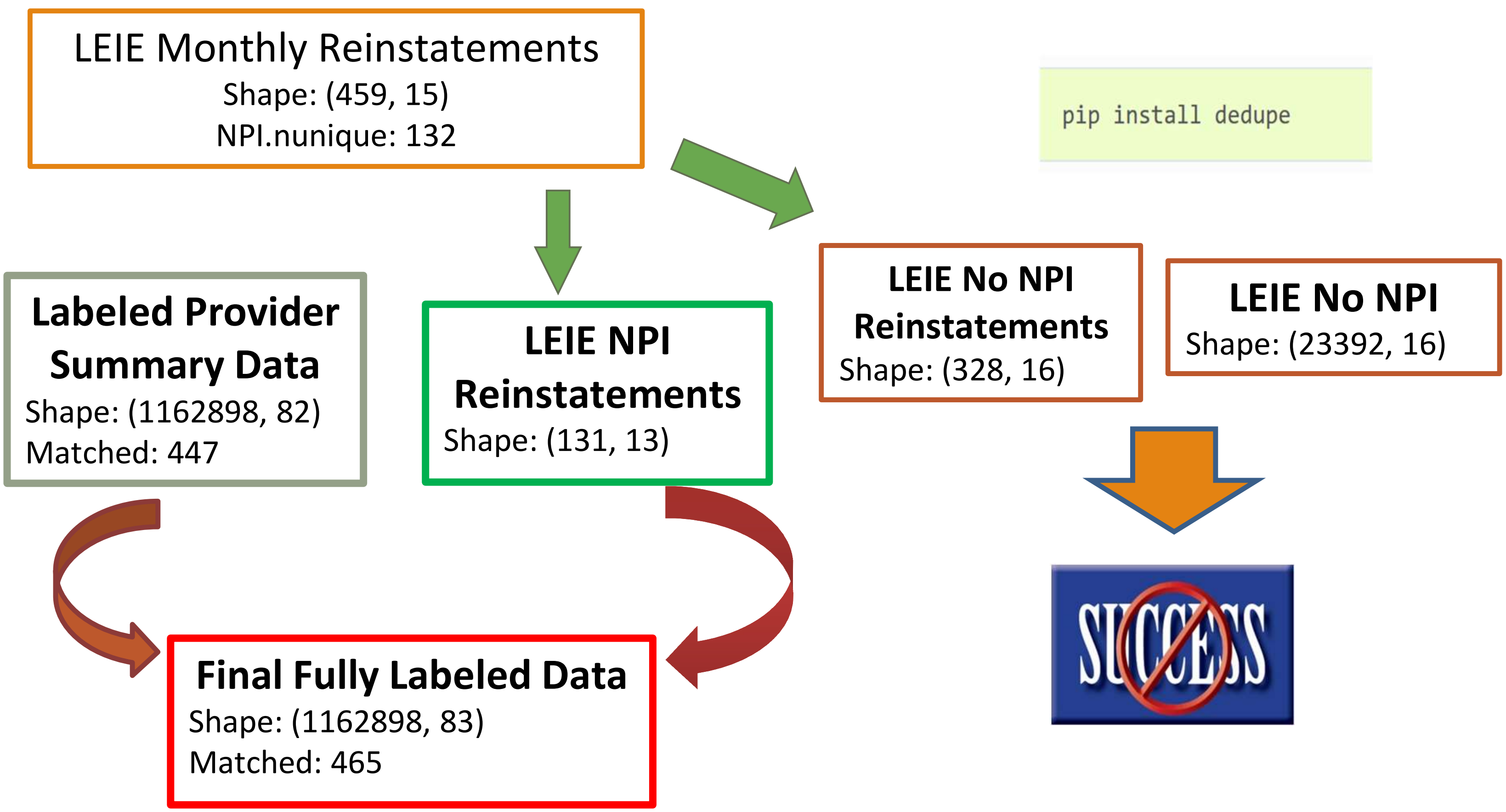
Filtering, Cleaning Values and Joining



Revision of Problem Statement:

Expand from fraud to all exclusions

As promised multiple times from the first day of class, we realized we needed more data…

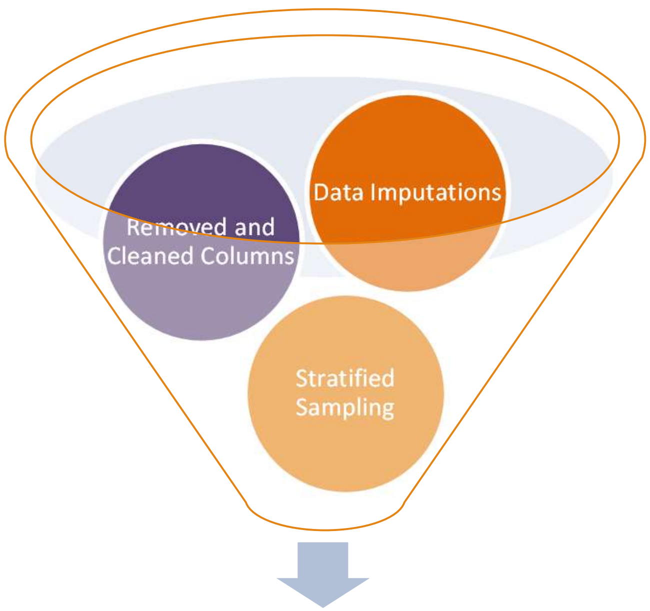


Final Wrangling

Fully Labeled Data

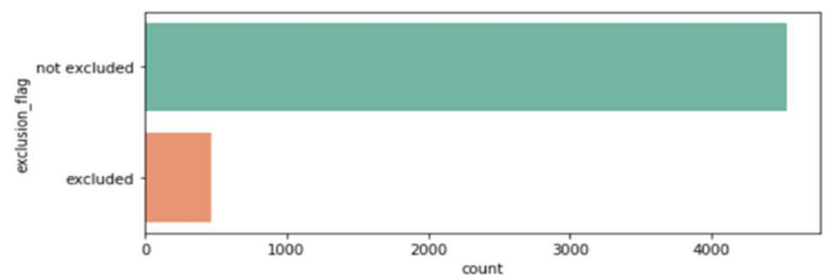
Shape: (1162898, 75)

Matched: 465



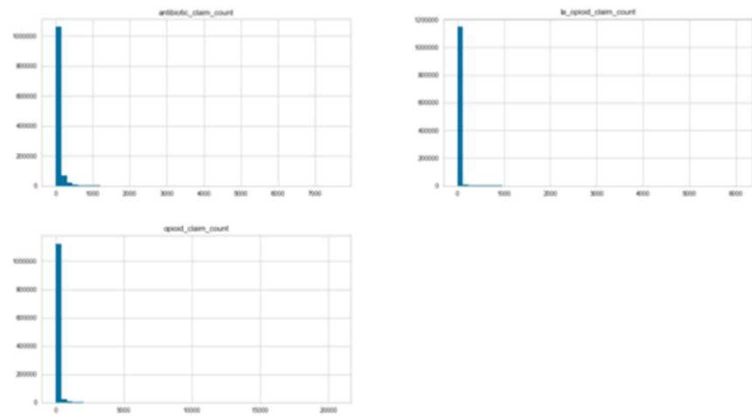
Final Dataset for Modeling

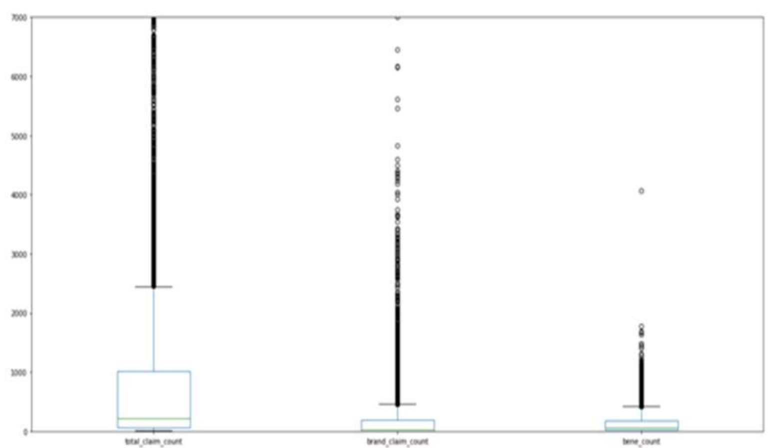
```
In [3]:  1 df.shape
Out[3]: (5000, 75)

In [4]:  1 df.columns
Out[4]: Index(['npi', 'nppes_provider_gender', 'specialty_description',
       'medicare_prvdr_enroll_status', 'total_claim_count',
       'total_30_day_fill_count', 'total_drug_cost', 'total_day_supply',
       'bene_count', 'ge65_suppress_flag', 'total_claim_count_ge65',
       'total_30_day_fill_count_ge65', 'total_drug_cost_ge65',
       'total_day_supply_ge65', 'bene_count_ge65_suppress_flag',
       'bene_count_ge65', 'brand_suppress_flag', 'brand_claim_count',
       'brand_drug_cost', 'generic_suppress_flag', 'generic_claim_count',
       'generic_drug_cost', 'other_suppress_flag', 'other_claim_count',
       'other_drug_cost', 'mapd_suppress_flag', 'mapd_claim_count',
       'mapd_drug_cost', 'pdp_suppress_flag', 'pdp_claim_count',
       'pdp_drug_cost', 'lis_suppress_flag', 'lis_claim_count',
       'lis_drug_cost', 'nonlis_suppress_flag', 'nonlis_claim_count',
       'nonlis_drug_cost', 'opioid_claim_count', 'opioid_drug_cost',
       'opioid_day_supply', 'opioid_bene_count', 'opioid_prescriber_rate',
       'la_opioid_claim_count', 'la_opioid_drug_cost', 'la_opioid_day_supply',
       'la_opioid_bene_count', 'la_opioid_prescriber_rate',
       'antibiotic_claim_count', 'antibiotic_drug_cost',
       'antibiotic_bene_count', 'antipsych_ge65_suppress_flag',
       'antipsych_claim_count_ge65', 'antipsych_drug_cost_ge65',
       'antipsych_bene_ge65_suppress_flg', 'antipsych_bene_count_ge65',
       'average_age_of_beneficiaries', 'beneficiary_female_count',
       'beneficiary_male_count', 'beneficiary_nondual_count',
       'beneficiary_dual_count', 'beneficiary_average_risk_score', 'EXCLYear',
       'REINYear', 'excl_type', 'exclusion_flag', 'country', 'state',
       'nppes_credentials', 'total_30_day_per_claim', 'drug_cost_per_claim',
       'day_supply_per_claim', 'female_count', 'male_count', 'nondual_count',
       'dual_count'],
      dtype='object')
```



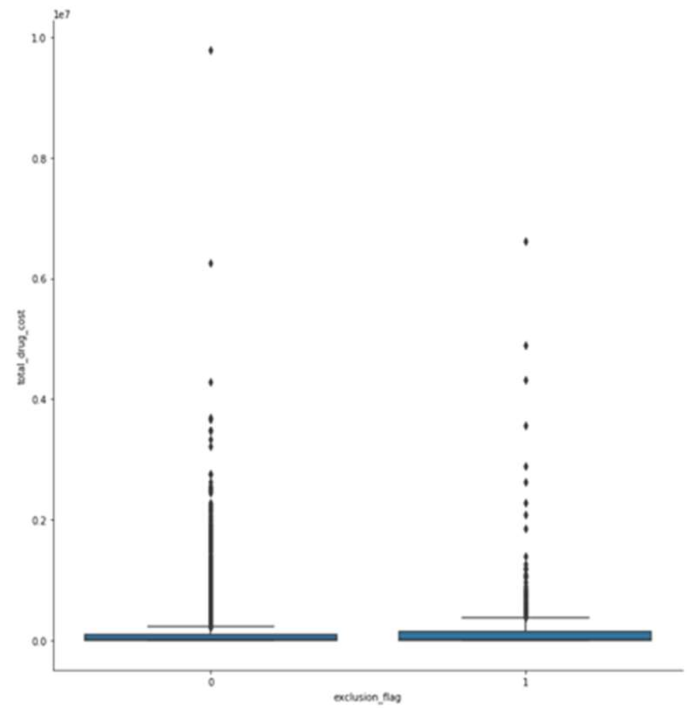
Feature Analysis And Selection

Data Distribution

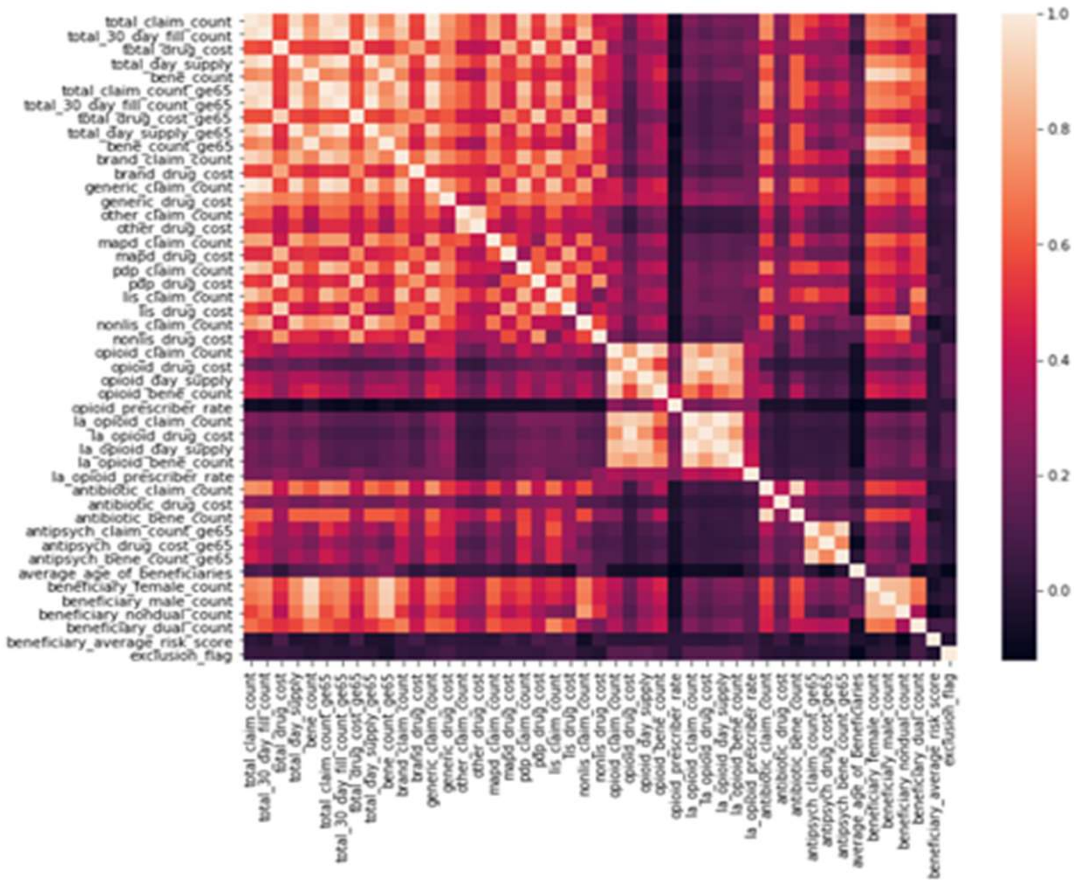




Data was skewed and will factor into our models



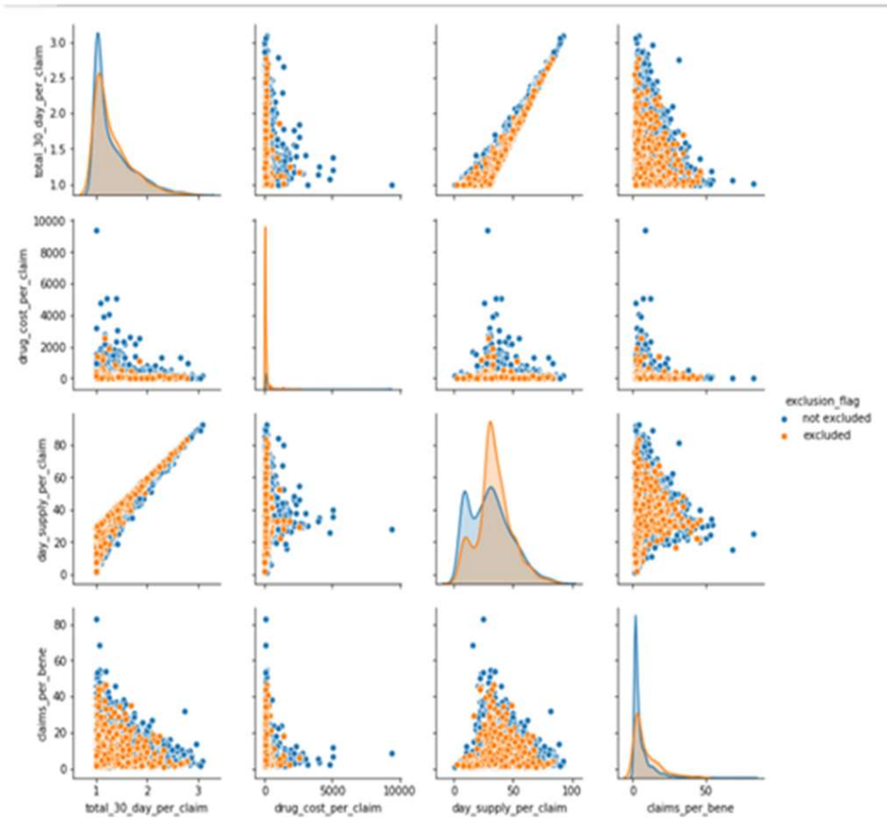
Correlation Matrix

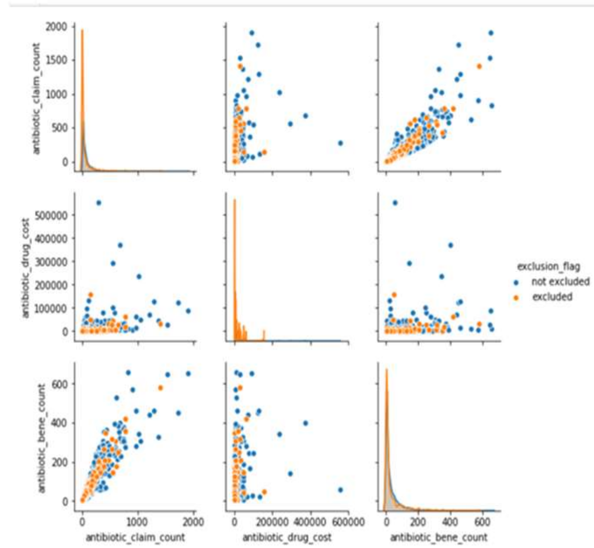


Noted high correlations on some features

Most made logical sense: total_claim_ count closely correlated with total_day_ supply

Pairplots And Targets

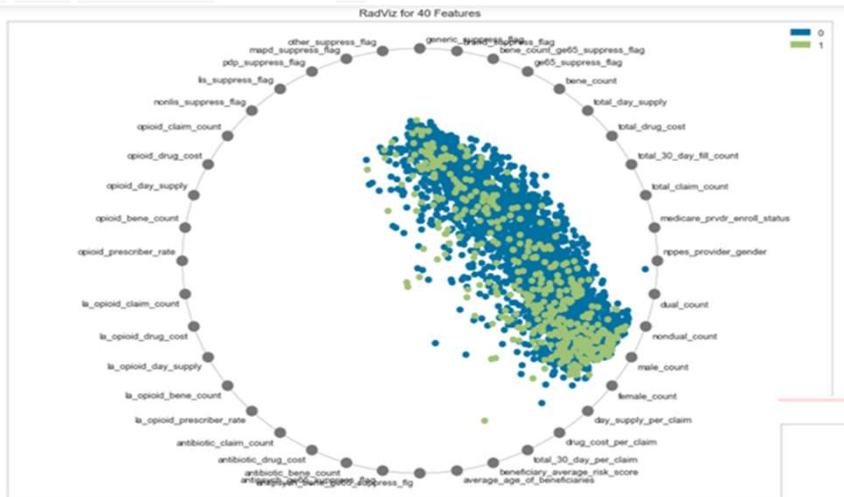




No obvious difference in clustering

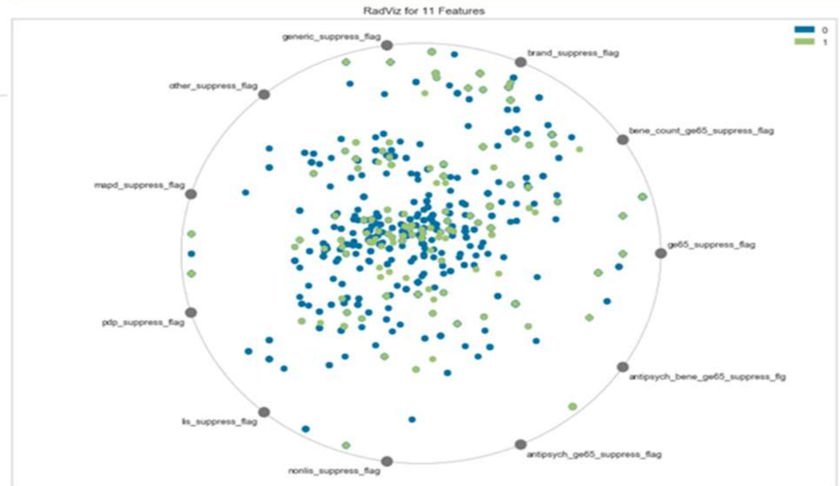
Overall, drug cost variables tended to be less closely correlated than other variables

Radial Visualization



Reviewed Radial visualization: split features by categorical variables for suppression of values and the columns of those values that may have been suppressed to see if one method showed different clustering than the other

No noticeable separability



Feature Selection

Transformer Methods

In [16]:
```
model = Lasso()
sfm = SelectFromModel(model)
sfm.fit(features, labels)
print(list(features.iloc[:, sfm.get_support(indices=True)]))
```

```
['bene_count_ge65', 'total_claim_count_ge65', 'lis_claim_count', 'opioid_claim_count']
```

In [17]:
```
model = Ridge()
sfm = SelectFromModel(model)
sfm.fit(features, labels)
print(list(features.iloc[:, sfm.get_support(indices=True)]))
```

```
['medicare_prvdr_enroll_status']
```

In [18]:
```
model = ElasticNet()
sfm = SelectFromModel(model)
sfm.fit(features, labels)
print(list(features.iloc[:, sfm.get_support(indices=True)]))
```

```
['bene_count_ge65', 'total_claim_count_ge65', 'opioid_claim_count', 'beneficiary_dual_count']
```

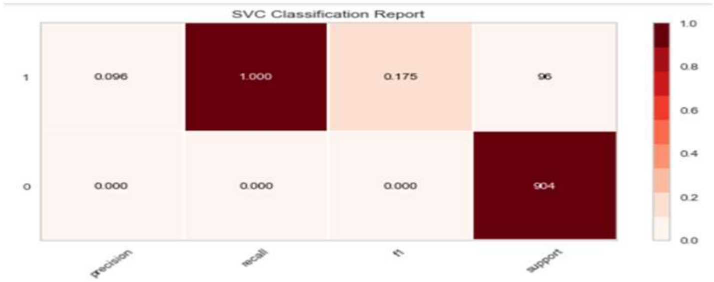
Relied on transformer methods and regularization techniques to decrease features included in the model as well as points learned from the literature and looking through the various visualizations of the features.

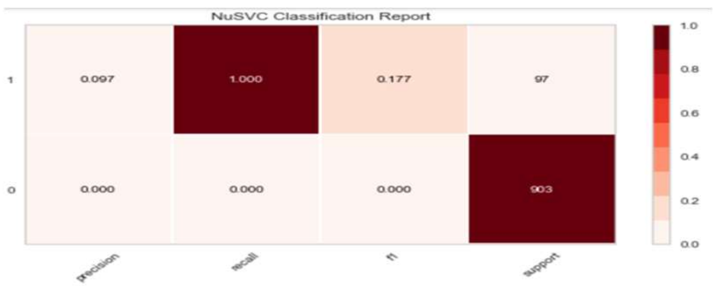
Models

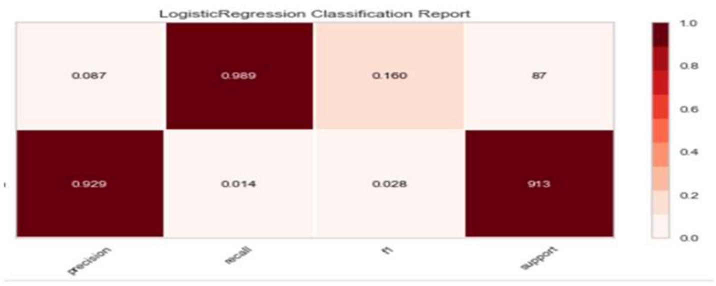
Initial Models

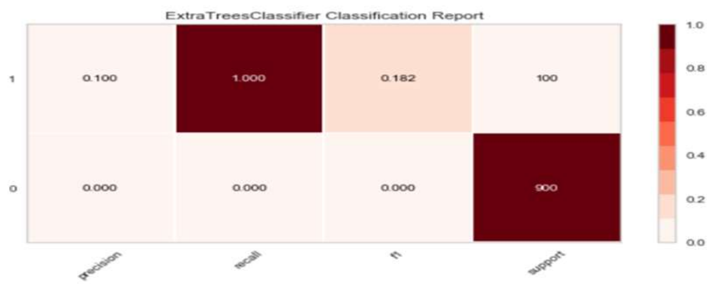
- Results show indication of class imbalance
- Model performance may be improved with standard scaler.
- Feature adjustments may also be made to improve performance.

```
SVC: 0.9359267734553776
NuSVC: 0.9757709251101321
KNeighborsClassifier: 0.15849056603773584
LinearSVC: 0.07647058823529412
SGDClassifier: 0.13917808219178082
LogisticRegression: 0.07243460764587525
LogisticRegressionCV: 0.056795131845841784
BaggingClassifier: 0.9468926553672317
ExtraTreesClassifier: 1.0
RandomForestClassifier: 1.0
```









Improving the Models

Filter down attributes that are the most likely predictive based on reading and data review

```
keep = ['medicare_prvdr_enroll_status', 'bene_count', 'bene_count_ge65', 'total_claim_count_ge65', 'brand_claim_count',
        'generic_drug_cost', 'other_drug_cost', 'mapd_drug_cost', 'pdp_claim_count', 'lis_claim_count',
        'nonlis_drug_cost', 'opioid_claim_count', 'la_opioid_bene_count', 'la_opioid_day_supply', 'opioid_prescriber_rate',
        'la_opioid_prescriber_rate', 'antipsych_drug_cost_ge65', 'antibiotic_drug_cost', 'average_age_of_beneficiaries',
        'beneficiary_dual_count', 'day_supply_per_claim', 'drug_cost_per_claim', 'exclusion_flag']
```

```
#Changing sample data
#Filtering to only the excluded rows
excluded = df_model1['exclusion_flag'] == 1
df_excluded = df_model1[excluded]
```

```
#Filtering to only non-excluded rows
not_excluded = df_model1['exclusion_flag'] == 0
df_not_excluded = df_model1[not_excluded]
df_not_excluded.shape
```

```
(4535, 23)
```

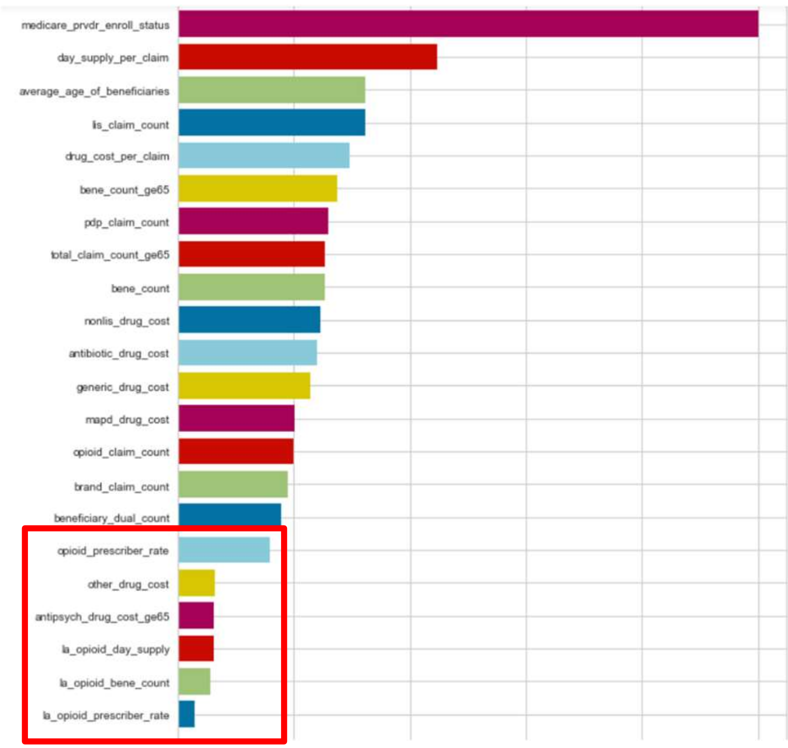
```
#Randomly samping the non excluded component
df_random = df_not_excluded.sample(n=500)
df_random.shape
```

```
(500, 23)
```

```
#Appending the randomly selected rows and the excluded rows
df_final = df_random.append(df_excluded, ignore_index = True)
```

Change the sample with an under sampling to fix class imbalances

Feature importance with Yellowbrick

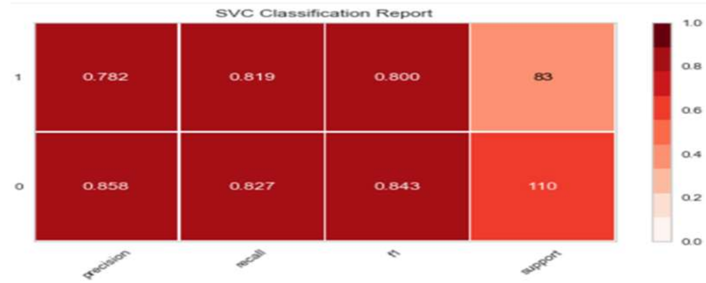


- Rank and plot relative importance of attributes
- Drop the bottom 5 features

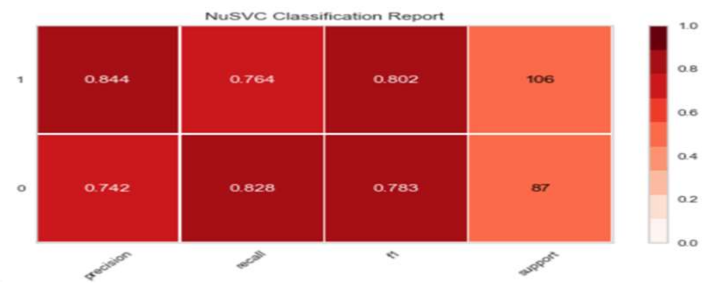
```
features = df_final.drop(columns = ['exclusion_flag', 'la_opioid_bene_count', 'la_opioid_day_supply',
                                    'other_drug_cost', 'antipsych_drug_cost_ge65', 'la_opioid_prescriber_rate']).columns
```

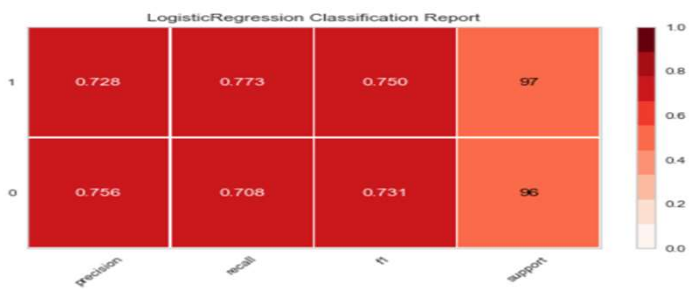
Final Models

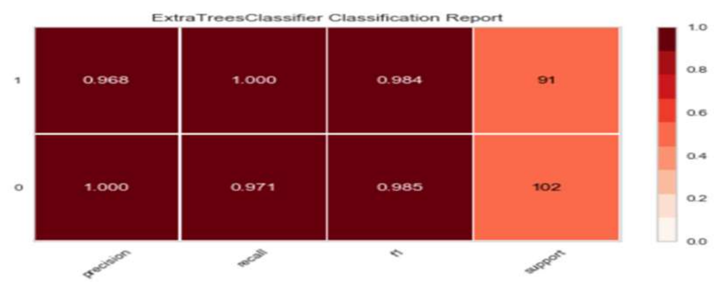
- Improvement for values in model classification reports.
- Notably increased values for F1 scores and recall.
- Reports also demonstrate balance.



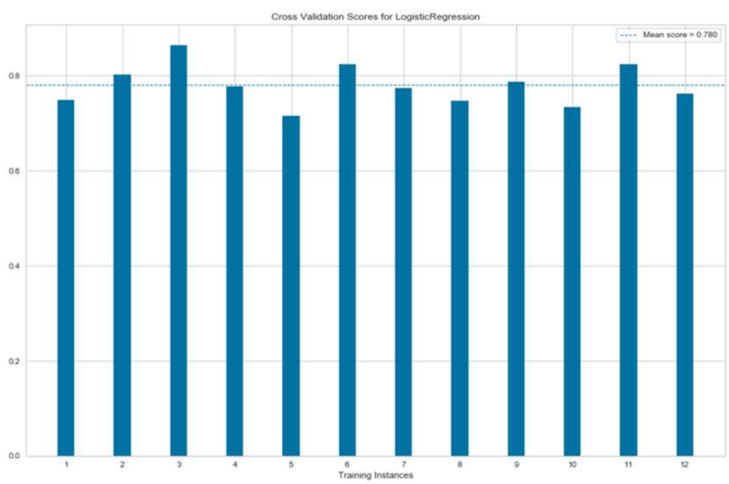

```
SVC: 0.783754116355653
NuSVC: 0.8362156663275687
KNeighborsClassifier: 0.8066298342541436
LinearSVC: 0.7524752475247525
SGDClassifier: 0.7193515704154002
LogisticRegression: 0.7505518763796911
LogisticRegressionCV: 0.7533039647577092
BaggingClassifier: 0.9738562091503269
ExtraTreesClassifier: 1.0
BalancedRandomForestClassifier: 0.9776833156216791
```







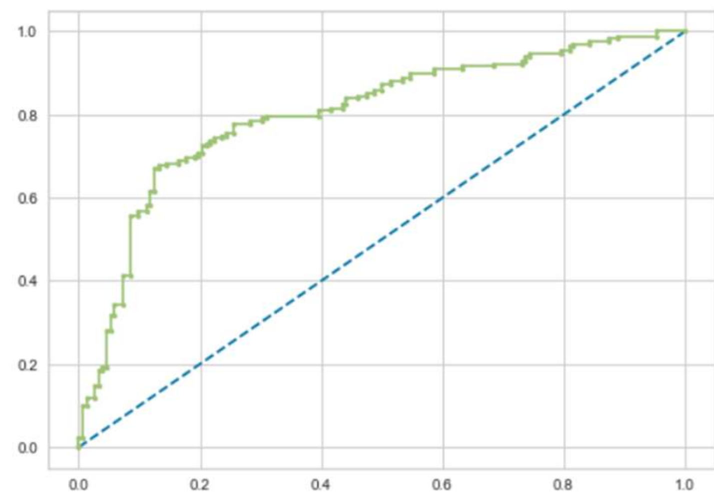
Final Model Evaluation Logistic Regression



Logistic Regression Cross Validation Mean Score of 0.780

Area Under Curve - Receiver (Operating Characteristics) curve

AUC Score: 0.80



Recommendations For Operationalizing The Model

- Next Steps:
  - Automate the monthly ingest process and analysis
  - Improve entity resolution and feature engineering to better fit model
  - Link Analysis to other Law Enforcement Databases – both Criminal and Civil

Recommendations For Value Creation

- Next Steps:
  - Improve provider identification
  - Reduce time to make criminal referrals on excluded entities and individuals
  - Identify Bad Actors to Improve Patient Safety and Oversight of Taxpayer Dollars

References

1. U.S. Department of Health and Human Services, Office of Inspector General, LEIE Downloadable Databases. https://oig.hhs.gov/exclusions/exclusions_list.asp#instruct
2. The Center for Medicare & Medicaid Services, Part D Prescriber Data FY 2017. https://www.cms.gov/index.php/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/PartD2017
3. The Center for Medicare & Medicaid Services, Part D Prescriber PUF Methodology. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Prescriber_Methods.pdf
4. Lematre G, Nogueira F, Aridas C.K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18(17), 1-5. Retrieved from http://jmlr.org/papers/v18/16-365

Thank You