Medicare fraud detection by using open source data

```python
class author(object):

    def __init__(self):

        self.name = "Xinyu (Max) Liu"

        self.email = "xinyulrsm@gmail.com"

        self.create =  "01/16/2016"

        self.addreess = "Waltham, MA 02453"
```

Outline

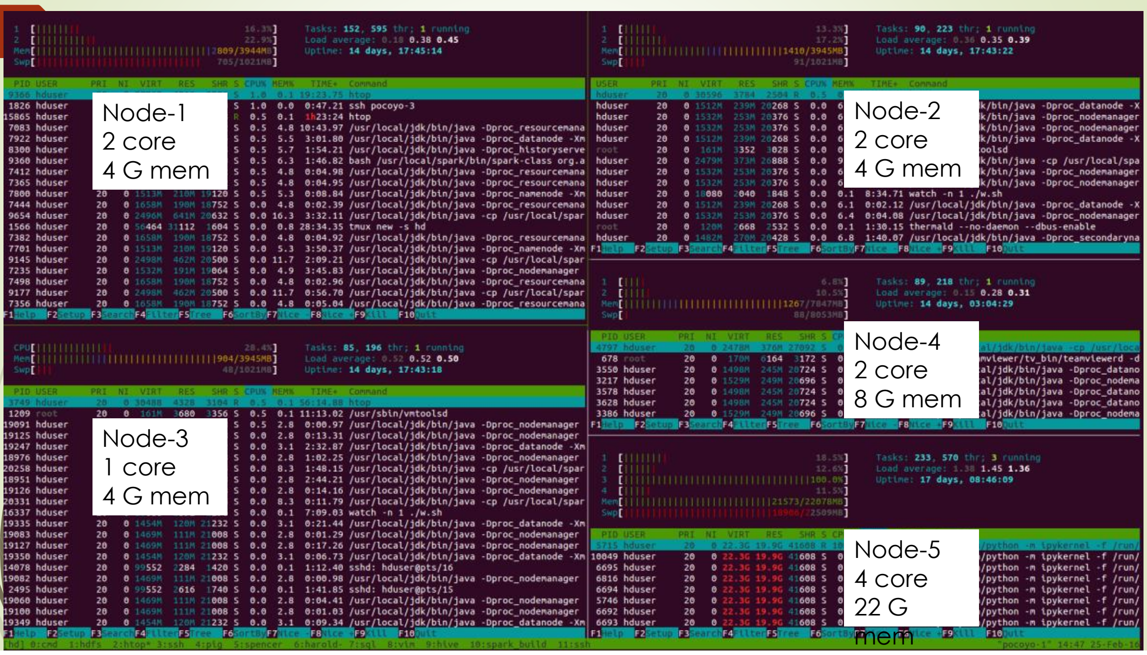
- Fraud detection is important in healthcare.
- The data source
- Feature engineering
- Models
- Results

Not many people but a lot of money got involved

Year	People Charged	False Billings, in Million
2010	94	$251
2011	91	$295
2012	2	$1.9
2013	89	$223
2014		
2015	243	$712

Data source: en.Wikipedia.org/wiki/Medicare_fraud

The cluster (5 nodes Hadoop/yarn/spark)



```
  1  [||||||||                       16.3%]     Tasks: 152, 595 thr; 1 running
  2  [||||||||||                     22.9%]     Load average: 0.18 0.38 0.45
  Mem[||||||||||||||||||||||||||||2809/3944MB]  Uptime: 14 days, 17:45:14
  Swp[||||||||||||||||||||||||||||  705/1021MB]

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
 9366 hduser                              S  1.0  0.1 19:23.75 htop
 1826 hduser                              S  1.0  0.0  0:47.21 ssh pocoyo-3
15865 hduser                              R  0.5  0.1  1h23:24 htop
 7083 hduser                              S  0.5  4.8 10:43.97 /usr/local/jdk/bin/java -Dproc_resourcemana
 7922 hduser                              S  0.5  5.5  3:01.80 /usr/local/jdk/bin/java -Dproc_datanode -Xm
 8300 hduser                              S  0.5  5.7  1:54.21 /usr/local/jdk/bin/java -Dproc_historyserve
 9360 hduser                              S  0.5  6.3  1:46.82 bash /usr/local/spark/bin/spark-class org.a
 7412 hduser                              S  0.5  4.8  0:04.98 /usr/local/jdk/bin/java -Dproc_resourcemana
 7365 hduser                              S  0.5  4.8  0:04.95 /usr/local/jdk/bin/java -Dproc_resourcemana
 7800 hduser     20   0 1513M  210M 19120 S  0.5  5.3  0:08.84 /usr/local/jdk/bin/java -Dproc_namenode -Xm
 7444 hduser     20   0 1658M  190M 18752 S  0.0  4.8  0:02.39 /usr/local/jdk/bin/java -Dproc_resourcemana
 9654 hduser     20   0 2496M  641M 20632 S  0.0 16.3  3:32.11 /usr/local/jdk/bin/java -cp /usr/local/spar
 1566 hduser     20   0 56464 31112  1604 S  0.0  0.8 28:34.35 tmux new -s hd
 7382 hduser     20   0 1658M  190M 18752 S  0.0  4.8  0:04.92 /usr/local/jdk/bin/java -Dproc_resourcemana
 7701 hduser     20   0 1513M  210M 19120 S  0.0  5.3  3:50.37 /usr/local/jdk/bin/java -Dproc_namenode -Xm
 9145 hduser     20   0 2498M  462M 20500 S  0.0 11.7  2:09.21 /usr/local/jdk/bin/java -cp /usr/local/spar
 7235 hduser     20   0 1532M  191M 19064 S  0.0  4.9  3:45.83 /usr/local/jdk/bin/java -Dproc_nodemanager
 7498 hduser     20   0 1658M  190M 18752 S  0.0  4.8  0:02.96 /usr/local/jdk/bin/java -Dproc_resourcemana
 9177 hduser     20   0 2498M  462M 20500 S  0.0 11.7  0:56.70 /usr/local/jdk/bin/java -cp /usr/local/spar
 7356 hduser     20   0 1658M  190M 18752 S  0.0  4.8  0:05.04 /usr/local/jdk/bin/java -Dproc_resourcemana
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

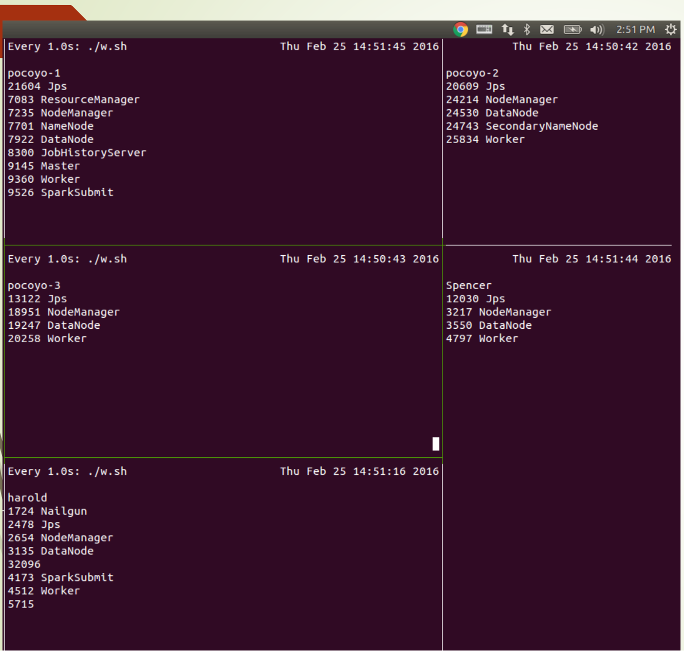
```
  CPU[|||||||||||||                  28.4%]     Tasks: 85, 196 thr; 1 running
  Mem[||||||||||||||||||||||||||||904/3945MB]   Load average: 0.52 0.52 0.50
  Swp[|||                            48/1021MB] Uptime: 14 days, 17:43:18

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
 3749 hduser     20   0 30488  4328  3104 R  0.5  0.1 56:14.88 htop
 1209 root       20   0  161M  3680  3356 S  0.5  0.1 11:13.02 /usr/sbin/vmtoolsd
19091 hduser                              S  0.5  2.8  0:00.97 /usr/local/jdk/bin/java -Dproc_nodemanager
19125 hduser                              S  0.0  2.8  0:13.31 /usr/local/jdk/bin/java -Dproc_nodemanager
19247 hduser                              S  0.0  3.1  2:32.87 /usr/local/jdk/bin/java -Dproc_datanode -Xm
18976 hduser                              S  0.0  2.8  1:02.25 /usr/local/jdk/bin/java -Dproc_nodemanager
20258 hduser                              S  0.0  8.3  1:48.15 /usr/local/jdk/bin/java -cp /usr/local/spar
18951 hduser                              S  0.0  2.8  2:44.21 /usr/local/jdk/bin/java -Dproc_nodemanager
19126 hduser                              S  0.0  2.8  0:14.16 /usr/local/jdk/bin/java -Dproc_nodemanager
20331 hduser                              S  0.0  8.3  0:11.79 /usr/local/jdk/bin/java -cp /usr/local/spar
16337 hduser                              S  0.0  0.1  7:09.03 watch -n 1 ./w.sh
19335 hduser     20   0 1454M  120M 21232 S  0.0  3.1  0:21.44 /usr/local/jdk/bin/java -Dproc_datanode -Xm
19083 hduser     20   0 1469M  111M 21008 S  0.0  2.8  0:01.29 /usr/local/jdk/bin/java -Dproc_nodemanager
19127 hduser     20   0 1469M  111M 21008 S  0.0  2.8  0:17.26 /usr/local/jdk/bin/java -Dproc_nodemanager
19350 hduser     20   0 1454M  120M 21232 S  0.0  3.1  0:06.73 /usr/local/jdk/bin/java -Dproc_datanode -Xm
14078 hduser     20   0 99552  2284  1420 S  0.0  0.1  1:12.40 sshd: hduser@pts/16
19082 hduser     20   0 1469M  111M 21008 S  0.0  2.8  0:00.98 /usr/local/jdk/bin/java -Dproc_nodemanager
 2495 hduser     20   0 99552  2616  1740 S  0.0  0.1  1:41.85 sshd: hduser@pts/15
19060 hduser     20   0 1469M  111M 21008 S  0.0  2.8  0:04.41 /usr/local/jdk/bin/java -Dproc_nodemanager
19100 hduser     20   0 1469M  111M 21008 S  0.0  2.8  0:01.03 /usr/local/jdk/bin/java -Dproc_nodemanager
19349 hduser     20   0 1454M  120M 21232 S  0.0  3.1  0:09.34 /usr/local/jdk/bin/java -Dproc_datanode -Xm
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
[hd] 0:cmd  1:hdfs  2:htop* 3:ssh  4:pig  5:spencer  6:harold- 7:sql  8:vim  9:hive  10:spark_build  11:ssh
```

```
  1  [||||||                         13.3%]     Tasks: 90, 223 thr; 1 running
  2  [||||||||                       17.2%]     Load average: 0.36 0.35 0.39
  Mem[|||||||||||||||||||||||||||1410/3945MB]   Uptime: 14 days, 17:43:22
  Swp[||||                           91/1021MB]

  USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
  hduser     20   0 30596  3784  2504 R  0.5  0
  hduser     20   0 1512M  239M 20268 S  0.0  6                  k/bin/java -Dproc_datanode -X
  hduser     20   0 1532M  253M 20376 S  0.0  6                  k/bin/java -Dproc_nodemanager
  hduser     20   0 1532M  253M 20376 S  0.0  6                  k/bin/java -Dproc_nodemanager
  hduser     20   0 1512M  239M 20268 S  0.0  6                  k/bin/java -Dproc_datanode -X
  root       20   0  161M  3352  3028 S  0.0  0                  oolsd
  hduser     20   0 2479M  373M 26888 S  0.0  9                  k/bin/java -cp /usr/local/spa
  hduser     20   0 1532M  253M 20376 S  0.0  6                  k/bin/java -Dproc_nodemanager
  hduser     20   0 1532M  253M 20376 S  0.0  6                  k/bin/java -Dproc_nodemanager
  hduser     20   0 18080  2040  1848 S  0.0  0.1  8:34.71 watch -n 1 ./w.sh
  hduser     20   0 1512M  239M 20268 S  0.0  6.1  0:02.12 /usr/local/jdk/bin/java -Dproc_datanode -X
  hduser     20   0 1532M  253M 20376 S  0.0  6.4  0:04.08 /usr/local/jdk/bin/java -Dproc_nodemanager
  root       20   0  120M  2668  2532 S  0.0  0.1  1:30.15 thermald --no-daemon --dbus-enable
  hduser     20   0 1482M  270M 20428 S  0.0  6.8  1:40.07 /usr/local/jdk/bin/java -Dproc_secondaryna
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

```
  1  [||||                            6.8%]     Tasks: 89, 218 thr; 1 running
  2  [|||||                          10.5%]     Load average: 0.15 0.28 0.31
  Mem[||||||||||||||||||||||||||1267/7847MB]    Uptime: 14 days, 03:04:29
  Swp[|                              88/8053MB]

  PID USER      PRI  NI  VIRT   RES   SHR S CP
 4797 hduser     20   0 2478M  376M 27092 S  0                al/jdk/bin/java -cp /usr/loca
  678 root       20   0  170M  6164  3172 S  0                mviewer/tv_bin/teamviewerd -d
 3550 hduser     20   0 1498M  245M 20724 S  0                al/jdk/bin/java -Dproc_datano
 3217 hduser     20   0 1529M  249M 20696 S  0                al/jdk/bin/java -Dproc_nodema
 3578 hduser     20   0 1498M  245M 20724 S  0                al/jdk/bin/java -Dproc_datano
 3628 hduser     20   0 1498M  245M 20724 S  0                al/jdk/bin/java -Dproc_datano
 3386 hduser     20   0 1529M  249M 20696 S  0                al/jdk/bin/java -Dproc_nodema
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

```
  1  [||||||||                       18.5%]     Tasks: 233, 570 thr; 3 running
  2  [||||||                         12.6%]     Load average: 1.38 1.45 1.36
  3  [|||||||||||||||||||||||||||||100.0%]     Uptime: 17 days, 08:46:09
  4  [|||||                          11.5%]
  Mem[|||||||||||||||||||||||||21573/22078MB]
  Swp[|||||||||||||||||||||||||18900/22509MB]

  PID USER      PRI  NI  VIRT   RES   SHR S CP
 5715 hduser     20   0 22.3G 19.9G 41608 R 10               /python -m ipykernel -f /run/
10049 hduser     20   0 22.3G 19.9G 41608 S  0               /python -m ipykernel -f /run/
 6695 hduser     20   0 22.3G 19.9G 41608 S  0               /python -m ipykernel -f /run/
 6816 hduser     20   0 22.3G 19.9G 41608 S  0               /python -m ipykernel -f /run/
 6694 hduser     20   0 22.3G 19.9G 41608 S  0               /python -m ipykernel -f /run/
 5746 hduser     20   0 22.3G 19.9G 41608 S  0               /python -m ipykernel -f /run/
 6692 hduser     20   0 22.3G 19.9G 41608 S  0               /python -m ipykernel -f /run/
 6693 hduser     20   0 22.3G 19.9G 41608 S  0               /python -m ipykernel -f /run/
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortBy            F9Kill  F10Quit
                                                          "pocoyo-1" 14:47 25-Feb-16
```

Another way to monitor the cluster



```
Every 1.0s: ./w.sh                              Thu Feb 25 14:51:45 2016

pocoyo-1
21604 Jps
7083 ResourceManager
7235 NodeManager
7701 NameNode
7922 DataNode
8300 JobHistoryServer
9145 Master
9360 Worker
9526 SparkSubmit
```

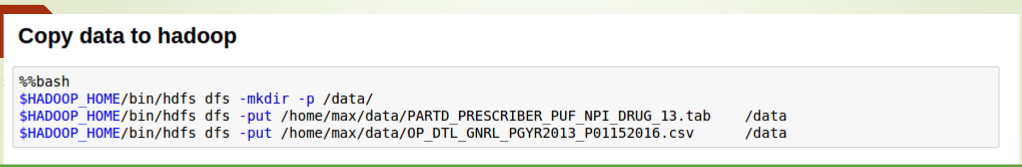
```
                                  Thu Feb 25 14:50:42 2016

pocoyo-2
20609 Jps
24214 NodeManager
24530 DataNode
24743 SecondaryNameNode
25834 Worker
```

```
Every 1.0s: ./w.sh                              Thu Feb 25 14:50:43 2016

pocoyo-3
13122 Jps
18951 NodeManager
19247 DataNode
20258 Worker
```

```
                                  Thu Feb 25 14:51:44 2016

Spencer
12030 Jps
3217 NodeManager
3550 DataNode
4797 Worker
```

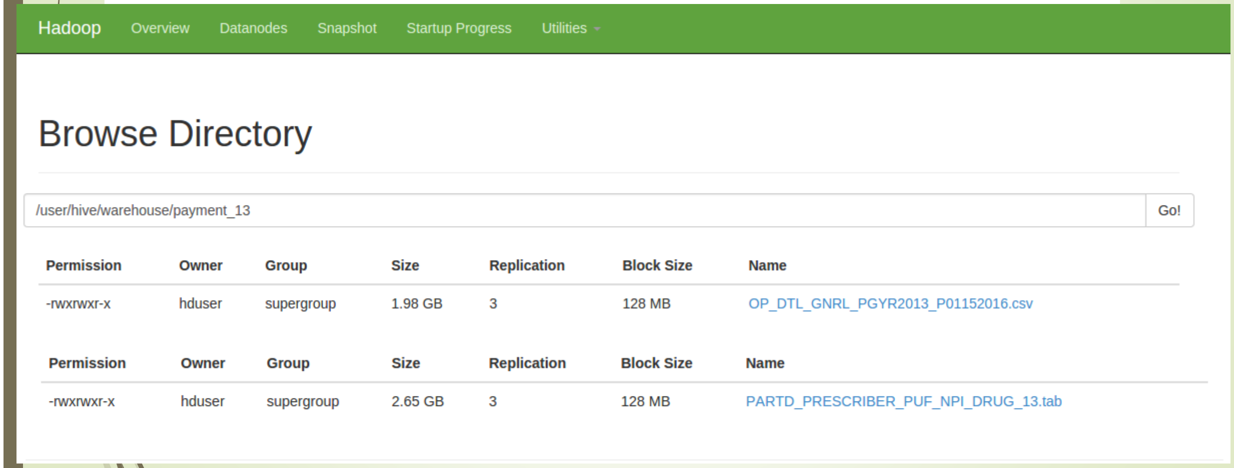
```
Every 1.0s: ./w.sh                              Thu Feb 25 14:51:16 2016

harold
1724 Nailgun
2478 Jps
2654 NodeManager
3135 DataNode
32096
4173 SparkSubmit
4512 Worker
5715
```

Copy files to Hadoop then move to Hive warehouse



Copy data to hadoop

```
%%bash
$HADOOP_HOME/bin/hdfs dfs -mkdir -p /data/
$HADOOP_HOME/bin/hdfs dfs -put /home/max/data/PARTD_PRESCRIBER_PUF_NPI_DRUG_13.tab    /data
$HADOOP_HOME/bin/hdfs dfs -put /home/max/data/OP_DTL_GNRL_PGYR2013_P01152016.csv      /data
```



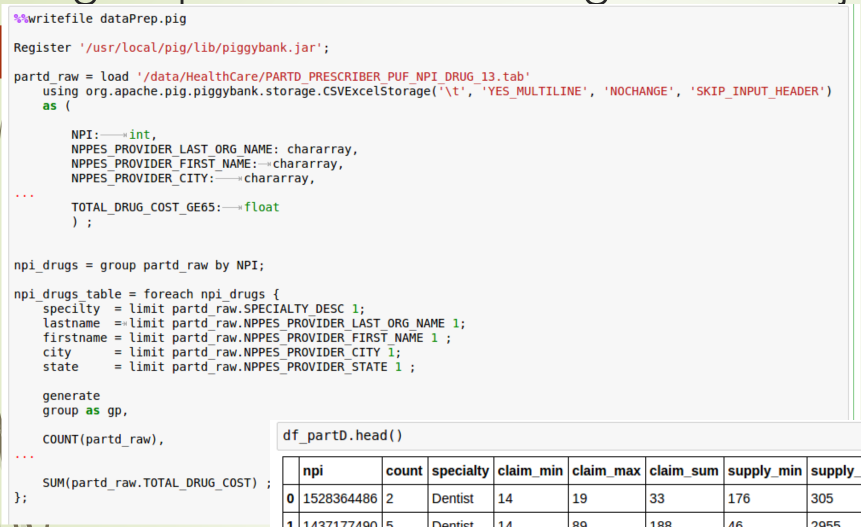
Browse Directory

/user/hive/warehouse/payment_13

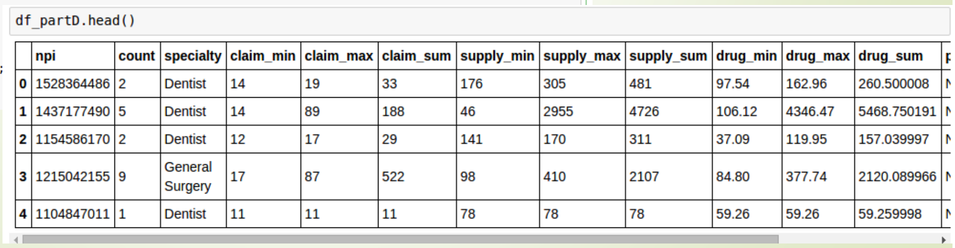
Permission	Owner	Group	Size	Replication	Block Size	Name
-rwxrwxr-x	hduser	supergroup	1.98 GB	3	128 MB	OP_DTL_GNRL_PGYR2013_P01152016.csv

Permission	Owner	Group	Size	Replication	Block Size	Name
-rwxrwxr-x	hduser	supergroup	2.65 GB	3	128 MB	PARTD_PRESCRIBER_PUF_NPI_DRUG_13.tab

Pig scripts for data cleaning and table join

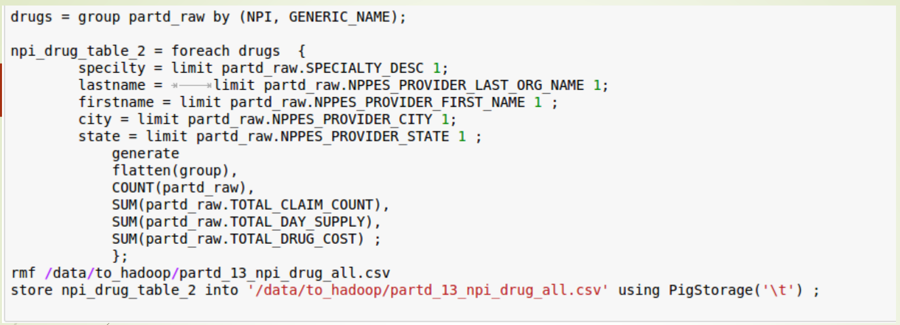


```
%%writefile dataPrep.pig

Register '/usr/local/pig/lib/piggybank.jar';

partd_raw = load '/data/HealthCare/PARTD_PRESCRIBER_PUF_NPI_DRUG_13.tab'
    using org.apache.pig.piggybank.storage.CSVExcelStorage('\t', 'YES_MULTILINE', 'NOCHANGE', 'SKIP_INPUT_HEADER')
    as (

        NPI:    int,
        NPPES_PROVIDER_LAST_ORG_NAME: chararray,
        NPPES_PROVIDER_FIRST_NAME:  chararray,
        NPPES_PROVIDER_CITY:    chararray,
...
        TOTAL_DRUG_COST_GE65:   float
        ) ;


npi_drugs = group partd_raw by NPI;

npi_drugs_table = foreach npi_drugs {
    specilty  = limit partd_raw.SPECIALTY_DESC 1;
    lastname  = limit partd_raw.NPPES_PROVIDER_LAST_ORG_NAME 1;
    firstname = limit partd_raw.NPPES_PROVIDER_FIRST_NAME 1 ;
    city      = limit partd_raw.NPPES_PROVIDER_CITY 1;
    state     = limit partd_raw.NPPES_PROVIDER_STATE 1 ;

    generate
    group as gp,

    COUNT(partd_raw),
...

    SUM(partd_raw.TOTAL_DRUG_COST) ;
};
```

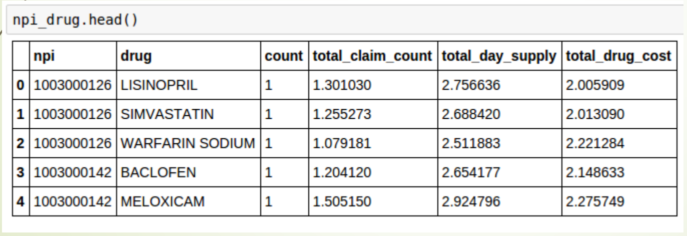


```
df_partD.head()
```

	npi	count	specialty	claim_min	claim_max	claim_sum	supply_min	supply_max	supply_sum	drug_min	drug_max	drug_sum	p
0	1528364486	2	Dentist	14	19	33	176	305	481	97.54	162.96	260.500008	N
1	1437177490	5	Dentist	14	89	188	46	2955	4726	106.12	4346.47	5468.750191	N
2	1154586170	2	Dentist	12	17	29	141	170	311	37.09	119.95	157.039997	N
3	1215042155	9	General Surgery	17	87	522	98	410	2107	84.80	377.74	2120.089966	N
4	1104847011	1	Dentist	11	11	11	78	78	78	59.26	59.26	59.259998	N

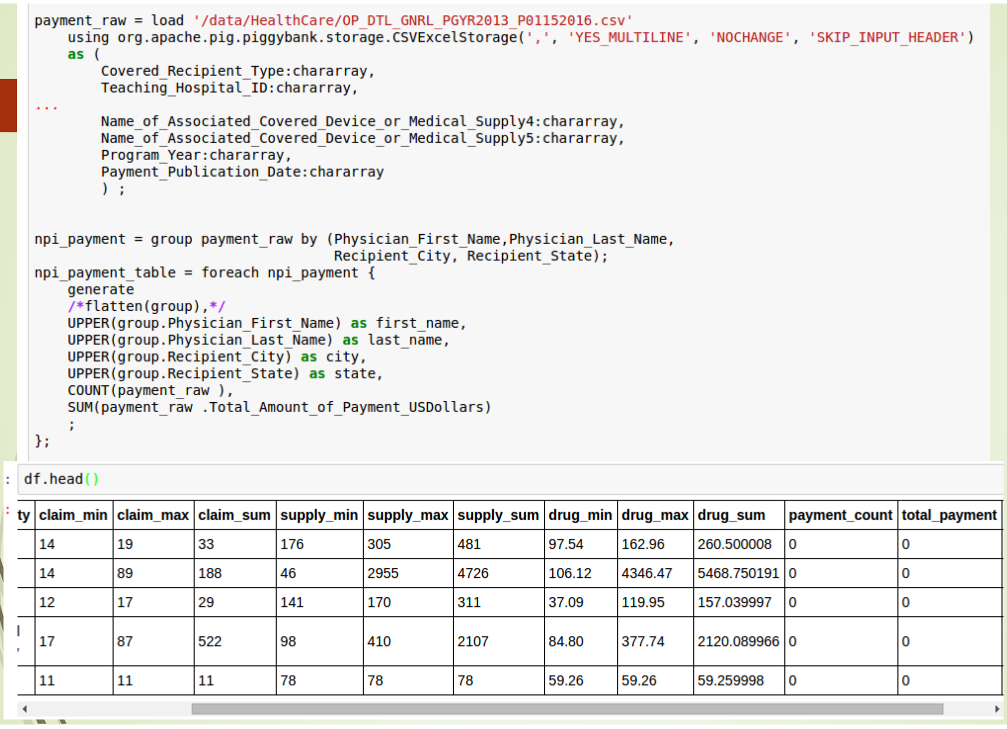


```
drugs = group partd_raw by (NPI, GENERIC_NAME);

npi_drug_table_2 = foreach drugs  {
        specilty = limit partd_raw.SPECIALTY_DESC 1;
        lastname =     limit partd_raw.NPPES_PROVIDER_LAST_ORG_NAME 1;
        firstname = limit partd_raw.NPPES_PROVIDER_FIRST_NAME 1 ;
        city = limit partd_raw.NPPES_PROVIDER_CITY 1;
        state = limit partd_raw.NPPES_PROVIDER_STATE 1 ;
            generate
            flatten(group),
            COUNT(partd_raw),
            SUM(partd_raw.TOTAL_CLAIM_COUNT),
            SUM(partd_raw.TOTAL_DAY_SUPPLY),
            SUM(partd_raw.TOTAL_DRUG_COST) ;
            };
rmf /data/to_hadoop/partd_13_npi_drug_all.csv
store npi_drug_table_2 into '/data/to_hadoop/partd_13_npi_drug_all.csv' using PigStorage('\t') ;
```



```
npi_drug.head()
```

	npi	drug	count	total_claim_count	total_day_supply	total_drug_cost
0	1003000126	LISINOPRIL	1	1.301030	2.756636	2.005909
1	1003000126	SIMVASTATIN	1	1.255273	2.688420	2.013090
2	1003000126	WARFARIN SODIUM	1	1.079181	2.511883	2.221284
3	1003000142	BACLOFEN	1	1.204120	2.654177	2.148633
4	1003000142	MELOXICAM	1	1.505150	2.924796	2.275749

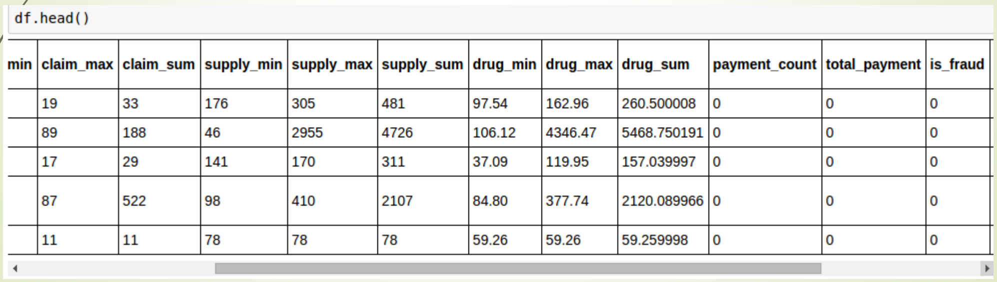


```
payment_raw = load '/data/HealthCare/OP_DTL_GNRL_PGYR2013_P01152016.csv'
    using org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE', 'NOCHANGE', 'SKIP_INPUT_HEADER')
    as (
        Covered_Recipient_Type:chararray,
        Teaching_Hospital_ID:chararray,
...
        Name_of_Associated_Covered_Device_or_Medical_Supply4:chararray,
        Name_of_Associated_Covered_Device_or_Medical_Supply5:chararray,
        Program_Year:chararray,
        Payment_Publication_Date:chararray
        ) ;


npi_payment = group payment_raw by (Physician_First_Name,Physician_Last_Name,
                                    Recipient_City, Recipient_State);
npi_payment_table = foreach npi_payment {
    generate
    /*flatten(group),*/
    UPPER(group.Physician_First_Name) as first_name,
    UPPER(group.Physician_Last_Name) as last_name,
    UPPER(group.Recipient_City) as city,
    UPPER(group.Recipient_State) as state,
    COUNT(payment_raw ),
    SUM(payment_raw .Total_Amount_of_Payment_USDollars)
    ;
};
```

```
df.head()
```

ty	claim_min	claim_max	claim_sum	supply_min	supply_max	supply_sum	drug_min	drug_max	drug_sum	payment_count	total_payment
	14	19	33	176	305	481	97.54	162.96	260.500008	0	0
	14	89	188	46	2955	4726	106.12	4346.47	5468.750191	0	0
	12	17	29	141	170	311	37.09	119.95	157.039997	0	0
l	17	87	522	98	410	2107	84.80	377.74	2120.089966	0	0
	11	11	11	78	78	78	59.26	59.26	59.259998	0	0

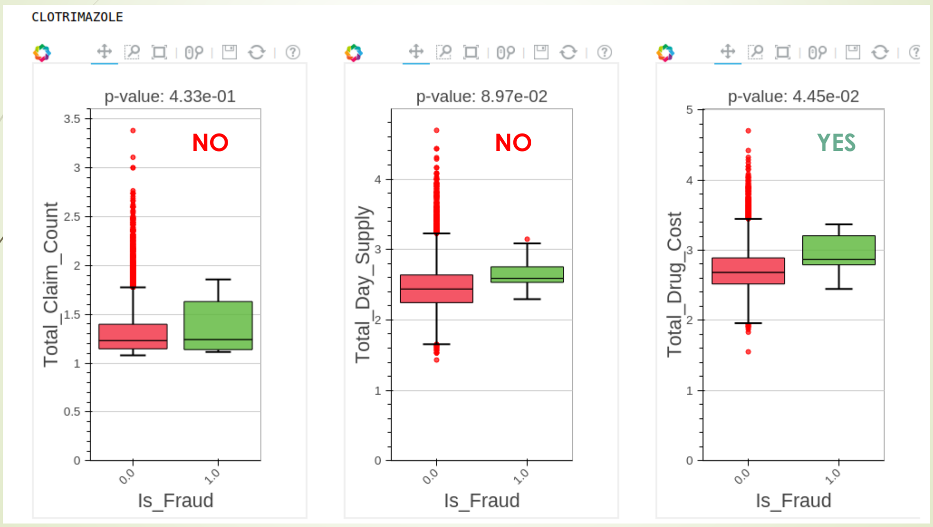
Npi exclusions database



```
df.head()
```

min	claim_max	claim_sum	supply_min	supply_max	supply_sum	drug_min	drug_max	drug_sum	payment_count	total_payment	is_fraud
	19	33	176	305	481	97.54	162.96	260.500008	0	0	0
	89	188	46	2955	4726	106.12	4346.47	5468.750191	0	0	0
	17	29	141	170	311	37.09	119.95	157.039997	0	0	0
	87	522	98	410	2107	84.80	377.74	2120.089966	0	0	0
	11	11	78	78	78	59.26	59.26	59.259998	0	0	0

An example
Select the feature with P<=0.05

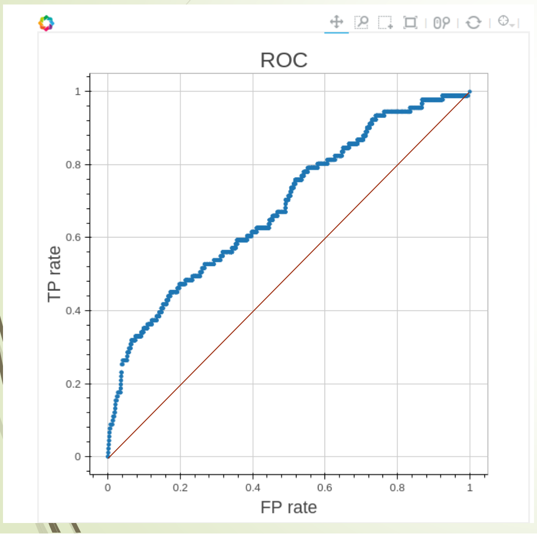


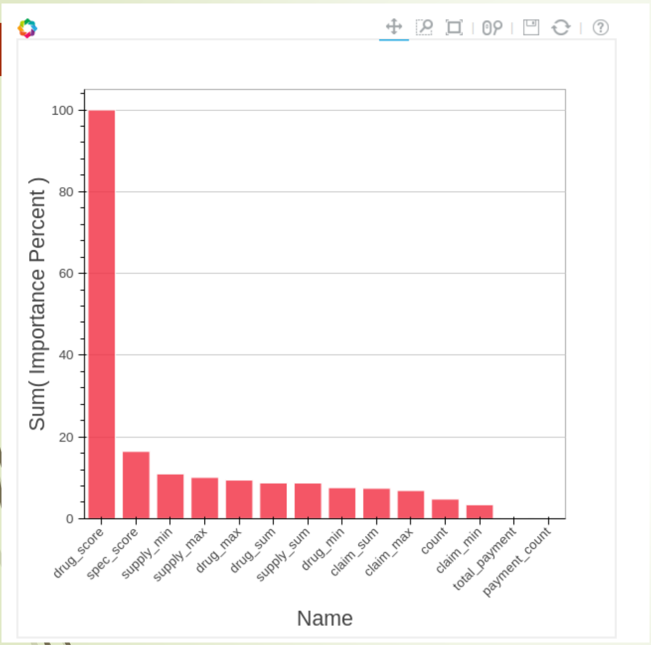
# Plots for the data, histogram or bar chart

Fitting results

AUC from RandomForest

AUC = 0.685





Ways to improve

- Use more data, for example the Part-B dataset.
- Add Page rank as a feature.
- Model blend might help.

References

- PageRank for Anomaly Detection by Ofer Mendelevitch and Jiwon Seo https://www.versustexas.com/criminal/healthcare-medicare-medicaid-fraud/