Project: Analyzing Public Health Data

Chalenge Masekera, Faye Ip, Ramit Malhotra

Info 290, Data-Mining (Spring 2014)
May 14, 2014

Introduction
Our objective in this project is to study patterns and outcomes of patient hospital visits at a Massachusetts hospital from 2006 to 2008. Health care reform in Massachusetts passed in 2006, and our team was curious to see what post-reform trends and patterns in hospital usage we could uncover, if any.

Related Studies
A previous study done using the same data set is “Emergency Department Utilization After the Implementation of Massachusetts Health Reform” by Smulowitz et al. The abstract can be found here: http://www.ncbi.nlm.nih.gov/pubmed/21570157. That study focused on comparing the use of the emergency room before and after health care reform passed in Massachusetts for conditions potentially amenable to primary care.

Data Cleaning
Our dataset is available here: https://www.dropbox.com/s/fawz0u3zkt4q9ow/HC_Reform.xls
Though our dataset was already in an Excel worksheet and in a next to usable state, we decided to make some transformations so that it better suited our needs. The first step involved merging the different Excel sheets into one csv file.

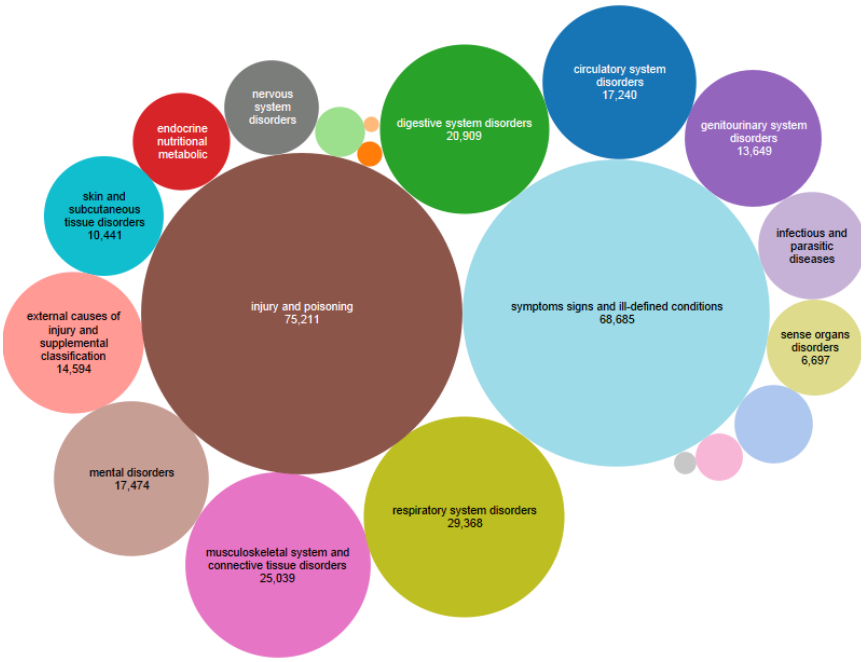
	A	B	C	D	E	F	G	H	I	J
1	ICD_9_CODE	DISPO	AGE	SEX	RACE	ZIP	Insur	INSURANCE	DOS	YEAR
34	274.0	1	43	M	Irish	01104	BLUE CROSS O MA-HM(	4	1-Oct-07	3
35	724.3	1	23	M	White	01071	BLUE CROSS O MA-HM(	4	1-Oct-07	3
36	913.0	1	15	M	White	01151	BLUE CROSS O MA-HM(	4	1-Oct-07	3
37	288.00	1	57	F	White	01060	BLUE CROSS O MA-HM(	4	1-Oct-07	3
38	789.00	1	35	F	White	06082	BLUE CROSS O MA-HM(	4	1-Oct-07	3
39	812.09	1	15	F	White	01085	BLUE CROSS O MA-HM(	4	1-Oct-07	3
40	786.2	1	32	M	White	01104	BLUE CROSS O MA-HM(	4	1-Oct-07	3
41	786.01	1	18	M	American	01056	BLUE CROSS O MA-HM(	4	1-Oct-07	3
42	789.00	1	31	F	White	01020	BLUE CROSS O MA-HM(	4	1-Oct-07	3
43	493.92	1	26	M	Puerto Rican	01104	BLUE CROSS- PPO	4	1-Oct-07	3
44	320.1	0	51	F	White	06082	BLUE CROSS- PPO	4	1-Oct-07	3
45	311	1	19	M	American	01104	BLUE CROSS- PPO	4	1-Oct-07	3
46	873.43	1	18	F	American	01028	BLUE CROSS- PPO	4	1-Oct-07	3
47	780.2	0	33	M	Unknown/Not Speci	01033	BLUE CROSS- PPO	4	1-Oct-07	3
48	784.0	1	18	F	White	01040	BLUE CROSS- PPO	4	1-Oct-07	3
49	847.0	1	39	M	African American	01104	BLUE CROSS- PPO	4	1-Oct-07	3
50	786.59	0	64	M	White	01095	BLUE CROSS- PPO	4	1-Oct-07	3
51	813.44	1	12	F	White	01028	BLUE CROSS- PPO	4	1-Oct-07	3
52	845.00	1	28	M	White	01077	BLUE CROSS- PPO	4	1-Oct-07	3
53	719.46	1	48	M	White	06103	BLUE CROSS- PPO	4	1-Oct-07	3
54	786.50	1	42	F	Unknown/Not Speci	01107	BLUE CROSS- PPO	4	1-Oct-07	3
55	845.00	1	18	M	White	01119	BLUE CROSS- PPO	4	1-Oct-07	3

The next step involved studying the data attributes. The initial data contained the following fields:

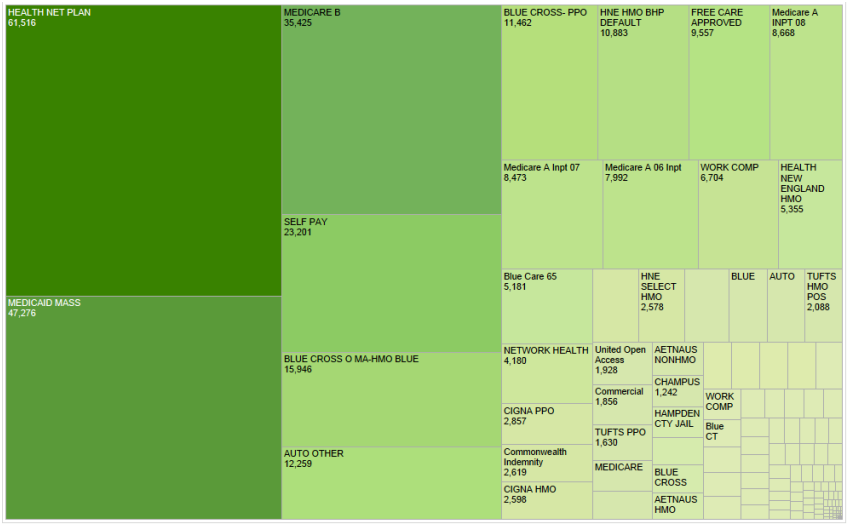
ICD_9_CODE - an alphanumeric field that is used to classify diseases and health related problems. This code is used by insurance companies to identify the specific diagnosis for which the patient received treatment.
DISPO - the numeric field that listed the outcome of a health visit, i.e. whether the patient got Discharged, Admitted or Eloped
AGE - continuous variable with the age of the patient
SEX - one letter character of either 'M/F' that listed the patients gender
RACE - the nationality of the patient as a string
ZIP - the zipcode of the patient's home address
INSURANCE - string with name of the patient's insurance company
INSUR - number representing a generic insurance policy category
DOS - date of service, i.e. the date when the patient was treated
YEAR - a number representing the year of treatment in ascending order with 2006 as 1.

The data in its original format was very granular and generated too many data points that made predictions very slow and difficult. We thus cleaned the data to represent a more usable format through field reclassification. For the ICD_9_CODE we grouped each code into the main disease classification as according to the established classifications. For example for codes between 290 and 319 we recoded them to category 'Mental disorder'. Similarly we changed the zip codes and replaced them with the counties in which the zip code belonged to. Lastly we transformed the outcomes to display the actual text instead of numbers.

Data Understanding
Our initial step involved getting to understand the data and what are properties of the data such as mean, standard deviation, number of unique variables. The next step involved some exploratory data analysis using Tableau to summarize the main characteristics of the data set such as which Insurance Company had the most clients, which disease category did most patients get treated for and the trends in the number of people that were being treated at the hospital over the timespan of the dataset.



This chart made in Tableau shows that the two largest diagnosis categories treated at the hospital are "injury and poisoning" and "symptoms, signs, and ill-defined conditions".



We also explored the relative sizes of the different insurance groups represented in the data set. This chart shows that about 1/3 of all visits are paid for by 3 insurance plans: Medicare B, Medicaid, and Health Net Plan, a managed care HMO group. Note that even though healthcare reform passed in 2006 requiring everyone to buy insurance, about 7% of patients are still “Self-Pay”.

Since all but one of the data fields are categorical, using SPSS, we plotted web graphs to map associations between different categories and discover likely patterns in the data set for example the association between the disease being treated and the likely outcome.

The above illustration gives an example of the associations we found, which shows ‘Self Pay’ insurance holders having a likely outcome of being Admitted. The most interesting discovery was how an overwhelming reated for external causes of injuries ended up ‘Eloping’. This important part of our research question allowed us to uncover different associations within our data set as well as likely impact on the algorithm we intended to implement.

Decision tree classification

The next step involved implementing a hospital outcome predictor that would predict what the likely treatment disposition is of a patient with specified characteristics - whether they would get Admitted, Discharged or they would Elope. Based on the features selected and we found the decision tree to be the most intuitive and easy to explain. Furthermore decision tree performance is usually not affected by nonlinear relationships between features. To implement this we used the decision tree algorithm in the scikit library for Python.

Since this is a supervised learning algorithm, we needed to have a training set and a test set. We split the data into two categories, i.e. the training which contained 99% of the data and the testing set which had the remaining data set.

1) Techniques and steps

- Preprocess: This involved initial cleaning of data and feature formats to convert the

strings to integers where possible.
- Prepare: This step vectorized nominal features and separated class labels (the 'disposition' feature' for training and testing.
- Train: We used 99.9% of the data to train our model, and cross validated it with 1% of the training data
- Test: We used the remainder of the dataset to test our model for accuracy!
- Output: The entire process led to this step which displayed the class predictions along with their accuracies and probabilities for each test row. We achieved a mean accuracy of around 72%.

Sample Output

```
Cross Validation Score
0.98
Mean Accuracy is:  0.725498883278
Classification Report:

             precision    recall  f1-score   support

          0       0.89      0.50      0.64     19867
          1       0.43      0.40      0.42     71271
          2       0.79      0.84      0.82    238398

avg / total       0.72      0.73      0.72    329536

Test Prediction Results
The top row is the actual record: ICD_9_CODE, DISEASE CATEGORY, DISPO, AGE, SEX, RACE, ZIP, COUNTY, INSURER, INSURANCE, DATE, YEAR.
The percentage-wise breakdown of the predicted disposition is below each of the actual records.
Record #0 : V64,external causes of injury and supplemental classification,Eloped,42,M,White,1105,Hampden County,SELF PAY,2,02-Oct-05,2006

['Eloped: 100.0%']

**************************
Record #1 : 784,symptoms signs and ill-defined conditions,Admitted,44,F,African American,1109,Hampden County,SELF PAY,2,02-Oct-05,2006

['Admitted: 100.0%']

**************************
Record #2 : 802,injury and poisoning,Admitted,36,M,White,12180,Essex County,SELF PAY,2,02-Oct-05,2006

['Admitted: 100.0%']

**************************
Record #3 : 786,symptoms signs and ill-defined conditions,Admitted,43,F,Puerto Rican,1151,Hampden County,SELF PAY,2,02-Oct-05,2006

['Admitted: 100.0%']

**************************
Record #4 : 625,genitourinary system disorders,Admitted,31,F,American,1057,Hampden County,SELF PAY,2,02-Oct-05,2006

['Admitted: 100.0%']

**************************
Record #5 : V64,external causes of injury and supplemental classification,Eloped,25,M,Irish,1082,Hampshire County,SELF PAY,2,02-Oct-05,2006

['Eloped: 100.0%']
```

2) Software Challenges

At first we ran into several problems due to our own limited understanding of decision trees and scikit-learn.

- Nominal categories: The first problem that was encountered was the inability of scikit to handle nominal features in dataset. It required changing all nominal categories to continuous data points - something that took us quite a while to figure out how to vectorize those feature sets.
- Overfitting: Next, we ran into a rather peculiar problem that at first didn't seem to suggest any issue. In our results, our model seemed to be predicting every row of data from our test set with a 100% confidence. We had a feeling that something was amiss, but we

couldn't really put our finger on it. After digging a little deeper (asking people in I School or on stackoverflow), we realized that we were doing two things incorrectly.

- Insufficient Training: We were using only 0.1% of the dataset for training our model. This was causing the model to overfit our data. To overcome this issue, we simply switched our test set and training set. Now our training set had 99.9% of the data, and our test set had 0.1%.
- Feature Selection: We were using too many features to train our model which also resulted in overfitting. The second problem was harder to resolve. We found 'Weka', a GUI software that allowed us to upload our dataset and perform analyses that included determination of the set of features that impacted classification the most.

Future work

- Validate using data from other hospitals or different years: We would like to validate our model using test data from other hospitals. If it continues to give us similar performance, it stands to reason that our the features which we have selected have a large correlation with the disposition outcomes of patients in hospitals.
- Anomaly Detection: For medical aid payments, it would be interesting to do some outlier analysis to sift through large hospital datasets in order to detect potential insurance fraud frequencies.
- Exploratory Mining: We would like to use other mining techniques in an exploratory fashion to find more patterns, to get a general understanding of what kind of mining works best for such datasets in order to discover interesting patterns that can be used for analysis.