Predicting Corporate Influence Cascades In Health Care Communities

Shouzhong Shi, Chaudary Zeeshan Arif, Sarah Tran

Stanford University

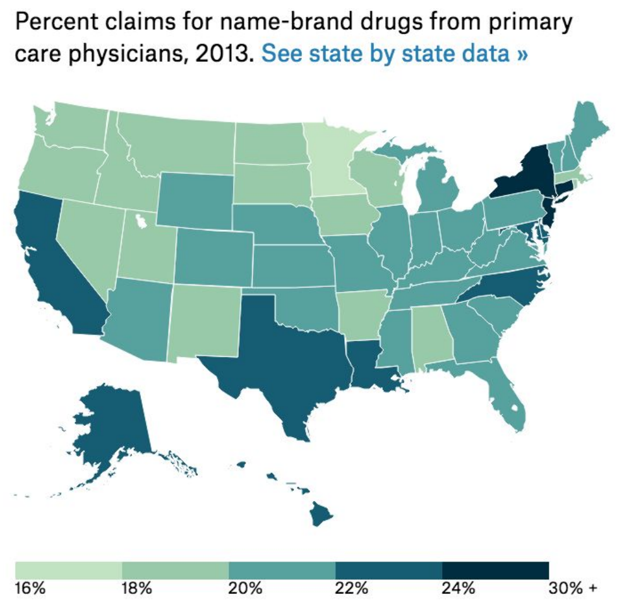
Data & Features

Question: Do the non-medical payments from manufacturers to physicians for specific drugs influence the choice of drugs the physicians prescribe?

Data Sources:

- **General Payment Data - Detailed Dataset 2013 Reporting Year**: 685,296 healthcare providers, 1,554 healthcare companies, and 2,098 teaching hospitals, 14.8 million general payment records, 825,947 research related payments and 9,768 ownership related payment, aggregated 3.6 million payment records.
- **Medicare Provider Utilization and Payment Data: 2013 Part D Prescriber:** 23,650,520 prescription types, 3,500 drug types

National Distribution of Prescription data [3]



Percent claims for name-brand drugs from primary care physicians, 2013. See state by state data »

Features: Filter payment records that have drug specified. Match the corresponding drug name with prescription record to establish the relationship of payment to prescription.

- Make 0, 1 vectors based on
- Same Physician practice or license state as the company business
- Range of payment amount such as >$100, >$500, >$1000, >$10000, >$25000, >$50000
- Drug claim count range
- Drug day supply range
- Drug Cost range
- Physician specialty vectors (500+ columns sparse).
- Physician's interest in ownership
- Nature of payment vectors such as gifts, charity, education, entertainment, food & beverage, travel & lodging etc.

Technology:

- Used 5-node Spark Cluster in AWS to perform feature extraction over the large dataset.
- PCA performed using the dataset using Spark APIs
- Used Weka for classifier algorithms.
- Used Matlab for classifier algorithms and to generate plots for bias and variance analysis.

Methods

Pre-Processing

- Filter the payment records which have associated drug names
- Join the Medicare Prescriptions on drug name match and matching physician first & last name. This establishes the positive case where the payment was received and the same drug was prescribed.
- Payment records where payment was received based on a drug association but was never prescribed is the negative case.
- Form vectors with all 0/1 features and train/test the dataset.

Logistic regression

The dependent variable in our datasets is limited (0 not prescribe / 1 prescribe), and logistic regression is a type of analysis where dependent variable is dummy variable.

The formula for binary logistic regression:

$P(Y=1| X= x) = 1 / (1 + e^{-(w0 + w1*x1 + w2*x2 + ...)}) = 1 / (1 + e^{-z})$

Log likelihood is:

$l(w) = (y_i * \log P(x_i; w) + (1-y_i) * \log(1-P(x_i,w)))$

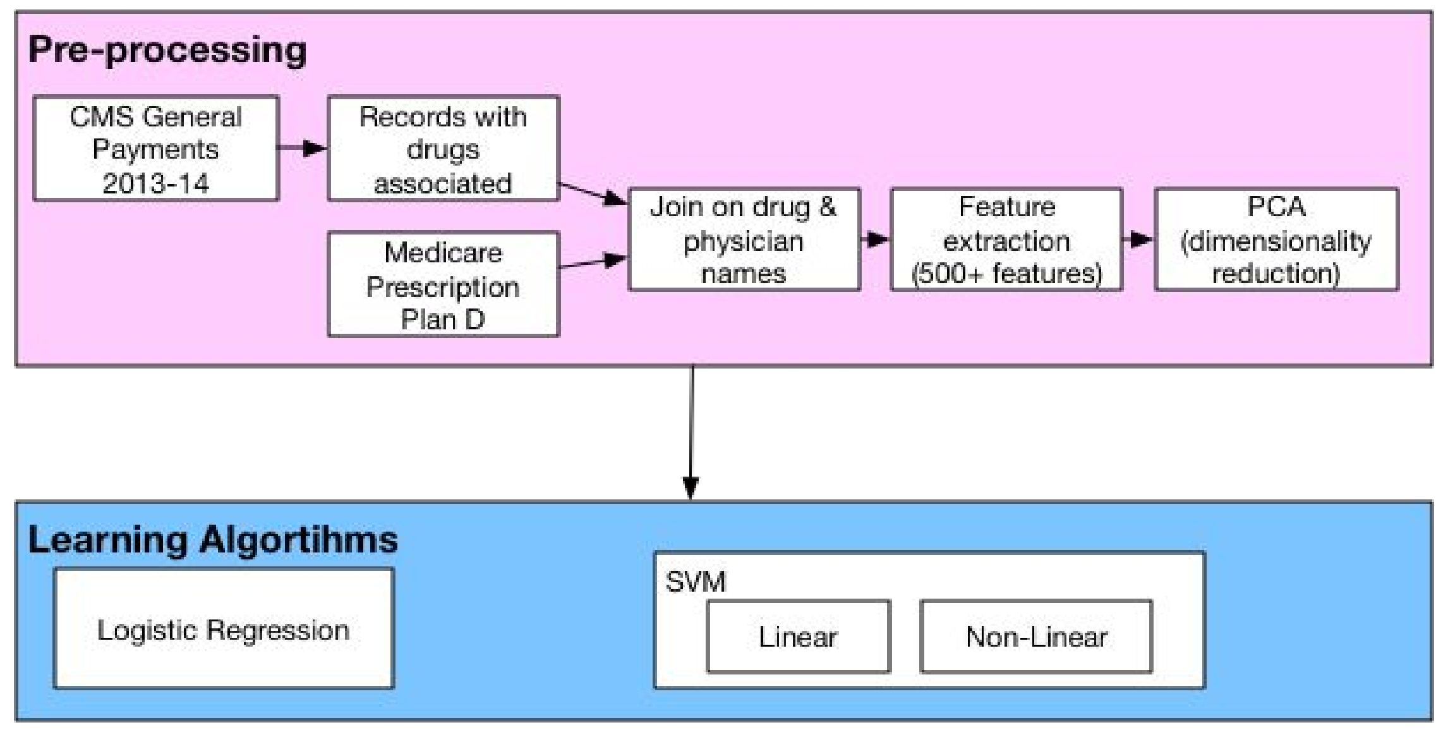
Using gradient ascent method can achieve the Maximum Likelihood Estimation.

Support Vection Machine

SVM is a state of the art classification algorithm. It is explicitly based on theoretical model of learning and has a lot of advantages, like not affected by local minima, does not suffer from the curse of dimensionality. It maximize the margin around the separating hyperplane, and it only specified by a subset of training samples, the support vectors. By adding the regularization term, it is quite tolerant to errors in prediction. Optimization for SVM is given as,

$min\ ||w||^2 + C * (\&)$ subject to, $y_i * (w\, t * X + b) >= 1 - \&$

Data Flow and Methodology Illustration



Pre-processing

CMS General Payments 2013-14

Records with drugs associated

Medicare Prescription Plan D

Join on drug & physician names

Feature extraction (500+ features)

PCA (dimensionality reduction)

Learning Algortihms

Logistic Regression

SVM

Linear

Non-Linear

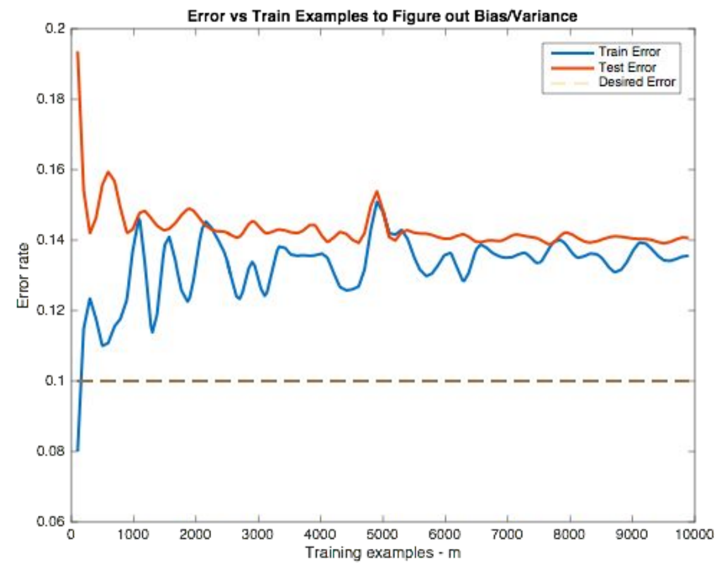
Results

Given the dataset and predefined features, we applied two classification algorithms and obtained the following results.

Logistic Regression Results:
Test Error (%): 13.95
Area Under PR: 0.859
Area Under ROC: 0.898

Support Vector Machines (SVM):
Test Error (%): 14.45
Area Under PR: 0.853
Area Under ROC: 0.853

Bias Variance Analysis:
Using SVM with a linear kernel achieved a 14% error rate. However, we wanted to see if we could make predictions with better accuracy. Based on our bias/variance analysis, we observed we had a bias problem.

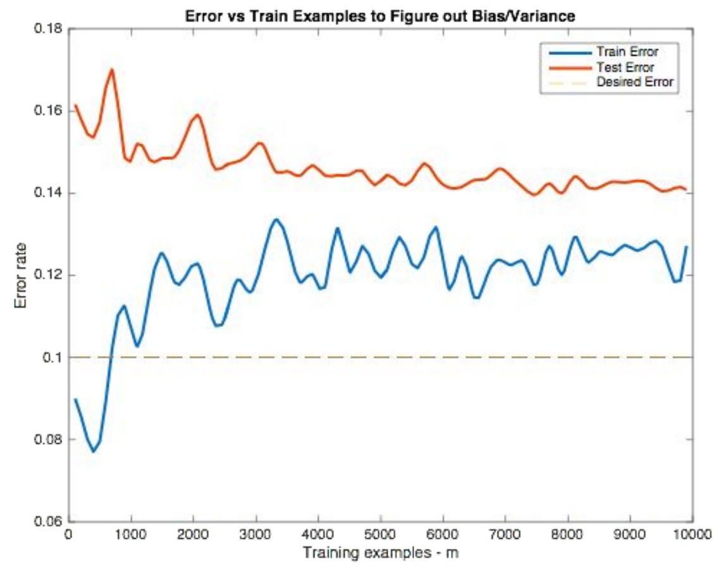


To address the bias issue, we tried

- Better modeling the dataset with a more complex model like a non-linear kernel to encapsulate the interactions between the features.
- Collecting more attributes for the dataset that would be more indicative of the likelihood of prescriptions.
- Improving the features set by removing noise by using PCA.

Using SVM with Polynomial Kernel of Degree 2:
Test Error (%): 14.09
Area Under PR: 0.856
Area Under ROC: 0.856

Although, the train set was better modeled with convergence of 12% error, using a nonlinear kernel did not improve the generalization/test error, still at ~14%. We now have a variance problem and are overfitting our training set.



Collecting more attributes for the dataset:
For selecting more features, we added a feature which indicates whether the physician practice or license state was the same as the company's business. This indirectly reflects the community geo-location information. The comparison of the two feature sets showed that with this feature, the accuracy improved by 3 percent for the training set.

Applying PCA to remove noisy features:
We used PCA and obtained the most meaningful 67 attributes from the feature set and removed the other noisy features. The 500+ column physician specialty features were found to be redundant and negatively impacting the hypothesis. Removing these resulted in improvement in reducing the test error.

Test Error (%): 13.47
Area Under PR: 0.862
Area Under ROC: 0.862

Conclusions

Dealing with two such large datasets, there was significant pre-processing and feature extraction. We orginally had many features, but not all the features were useful. Using PCA, we removed some noisy features.

To test our hypothesis, we used two classification algorithms, logistic regression and SVM. These two algorithms achieved similar results in our experiments using a linear model, although SVM takes more computation time. For further analysis, we used the SVM model.

The experimental results indicate the our original hypothesis holds. Non-medical payments seemingly have an influence in the physician community, especially based on geo-location information. However, with a 14% error rate, it is difficult to definitively say this.

In order to gain better prediction quality and to mitigate our bias issue, we will need to collect more features possibly in the form of detailed drug facts and patient symptom records. We believe physicians tend to recommend well-established and known drugs more often. Also, we think patient symptoms and preferences can play a significant role in prescription decision. These facts are currently not represented in our dataset.

Finally, based on our bias/variance analysis, our models represents the current dataset well and using a more complex model will cause overfitting for training data.

Bibliography:

- https://openpaymentsdata.cms.gov/
- https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html
- http://projects.propublica.org/checkup/