Machine Learning from Health Insurance Administrative Data: Opioids, Obamacare, and Other Applications

Kush R. Varshney
krvarshn@us.ibm.com
http://krvarshney.github.io

Machine Learning on Administrative Data: Lots of Examples I’ve Worked On

- HR and compensation data to predict voluntary resignation of IBM employees
- Work products data to estimate skills and expertise of IBM employees
- Enterprise data to predict successful team compositions of IBM employees
- Roll call, bill co-sponsorship, and bill text data to predict voting patterns of members of Congress
- Application form and repayment data to predict behaviors of pay-as-you-go solar power customers in rural India
- Application form data and evaluation history to predict recipients of a prestigious social entrepreneurship fellowship
- Grand slam tennis match statistics data to predict the winner of each point
Medical Claims

Used by payers (insurance companies) to reimburse health providers (physicians, hospitals, etc.)

Date of service

Diagnosis codes (ICD-10), procedure codes (CPT), drug codes (NDC)

Billed amount and paid amount

Provider and patient information (demographics)

Useful for many different machine learning tasks
  - Cost prediction, e.g. Obamacare
  - Understanding health patterns, e.g. opioid addiction

A Tale of Two Laws

PATIENT PROTECTION AND AFFORDABLE CARE ACT

Changed the landscape of the health insurance market in the United States

Health insurance companies had to decide which new markets to enter
  - New markets defined by geography, age group, and other prospect base criteria

HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT

Required all releases of health-related information about individuals to protect their privacy
  - Even for health insurance companies’ internal planning uses

k-Anonymity is a common mathematical interpretation of the privacy condition
Desiderata of Health Insurance Companies

Desire low-cost (healthy) people enroll in their plans
Not allowed to accept or deny enrollment on an individual basis
Allowed to offer or not offer plans in well-defined markets
(Allowed to have marketing strategies)
Use data-driven decision making for determining whether or not to offer plans in new markets

Desiderata of Health Insurance Companies

Desire cost data on people who will enroll in new markets
Only have cost data on people who have enrolled in existing markets
Have demographic data on existing market
Have demographic data on new market
Regression problem with covariate shift
  – Also need to consider enrollment: three-population shift
Demographic and Cost Data Availability

<table>
<thead>
<tr>
<th></th>
<th>Existing Market</th>
<th>New Market</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrolled</td>
<td>insurance company has demographic data and cost data</td>
<td>insurance company has no demographic data or cost data</td>
</tr>
<tr>
<td>Everyone (enrolled and not enrolled)</td>
<td>insurance company can get demographic data from public sources</td>
<td>insurance company can get demographic data from public sources</td>
</tr>
</tbody>
</table>

The Market Risk Assessment Regression Problem

Use cost and demographic data for enrolled members in the existing market, demographic data for the existing market, and demographic data for the new market to estimate cost for enrolled members in the new market.

Can use regression technique of choice
- Ordinary least-squares with and without log-transformed data
- Two-part models
- Generalized linear models
- Multiplicative regression

Need a type of covariate shift to account for the differences between the features of the existing and new market

For the three population shift and regression workload, need a privacy transformation that preserves the probability distribution of the data
Results

Developed such an approach, which required a new privacy-preservation method

Excellent empirical performance

Successfully used by a large health insurance company


Best Research Paper Honorable Mention

Opioid Epidemic

175 deaths per day due to opioid overdose in the United States

Physicians could be indirectly contributing to the epidemic by overprescription

One reason is because they have generally lacked data to guide opioid prescribing decisions

Use medical claims data to explore the causal relationships between the characteristics of the initial opioid prescriptions and outcomes
Results

In patients who are given synthetic opioids for no more than 7 days, using natural or semi-synthetic opioids instead could potentially reduce the risk of long-term use or addiction by 36.5%.

For natural or semi-synthetic opioids, a shorter days of supply could potentially reduce the risk by 65.3% in patients with longer days of supply.

Longer days of supply on the initial opioid prescription is a driving force of long-term use

A significant difference in average treatment effect between synthetic opioids and natural or semi-synthetic ones
