2014 CCP: Data Scientist – Medicare Claim Anomaly Detection Challenge

Welcome to the 2014 CCP: Data Scientist – Medicare Claim Anomaly Detection Challenge! This site will give you everything you need to get started.

Challenge Background

The US healthcare system has featured prominently in the news recently. Medicare, a national social health insurance program administered by the U.S. federal government, continues to be under pressure to ensure that its funds are spent efficiently. In the Medicare system, private providers submit claims to Medicare for medical procedures performed for covered individuals. Those claims are then evaluated for payment, with Medicare typically reimbursing about half of the procedures? costs on average. Because of pressures and budgetary constraints, the Medicare program needs to ensure that the type of procedures performed and their costs, are both consistent and reasonable. In addition, any errors or fraudulent claims from providers should be discovered.

You have been contracted to analyze a set of claims data from the Medicare program. Your job is to detect abnormal data – providers, areas, and patients with unusual procedures and/or claims.

Challenge Data

You have access to the following summary data, which aggregates information on procedures performed and billed by providers in 2011, as well as how much Medicare reimbursed:

- https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient.html
- https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Outpatient.html

These data include both inpatient and outpatient procedures. Inpatient procedures are coded with DRG codes (https://en.wikipedia.org/wiki/Diagnosis-related_group), and outpatient procedures with APC codes (https://en.wikipedia.org/wiki/Ambulatory_Payment_Classification). Note that a single provider may submit claims for both inpatient and outpatient procedures, and a single patient may have undergone inpatient and outpatient procedures during the same period.

In addition to the summary data, you also have access to individual patient procedure claims from 2013 and account records for individual patients. (NOTE: All data provided except for the summary inpatient and outpatient claims files are artificial and do not represent actual patient data.) The patient record data is presented as an XML dump from the Medicare patient database. All patient records have been completely anonymized and contain only a unique patient ID number and demographic information. The procedure claim records contain the procedures claimed by patient ID and date as ASCII-delimited text. (Seehttps://en.wikipedia.org/wiki/Delimiter#ASCII_delimited_text.) The date format is as prescribed by the ASC X12 837 Professional Health Care Claim Transaction, i.e. CCYYMMDD.
(See http://www.bcbsnc.com/assets/providers/public/pdfs/837Institutional_5010_v2.5.pdf.)

The patient and procedure data files are available here:

- https://cloudera.box.com/s/7h2sx81w73mm1ciswt5f

Challenge Problems

You have been contracted to analyze a set of claims data from the Medicare program for abnormalities. Specifically, you've been asked to answer the following three queries:

1. Some providers and regions may be consistently billing too much for procedures or billing for too many procedures, perhaps inadvertently. To better focus investigative efforts, the staff has asked you to answer the following questions:
    - Which **three** procedures have the highest relative variance in cost?
    - Which **three** providers claimed the highest amount (on average) for the largest number of procedures?
      To clarify, consider this example. If out of four providers, Provider A had the highest average claim amounts for procedures 1, 2, and 3, Provider B had the highest average claim amount for procedure 4, and Provider C and Provider D did not have the highest average claim amount for any procedure, then Provider A claimed the highest amount for the largest number of procedures (3 versus 1, 0, and 0).
    - The providers in which **three** regions claimed the highest average amount for the largest number of procedures? The region can be found in the *Hospital Referral Region Description* column in the summary data files.
      To clarify, consider this example. If out of three regions, providers in San Jose, CA had the highest average claim amount for procedure 1 (averaged across all providers in San Jose), providers in Boston, MA had the highest average claim amounts for procedures 2 and 3 (averaged across all providers in Boston), and providers in New York, NY had the highest average claim amount for procedure 4 (averaged across all providers in New York), then the providers in Boston, MA claimed the highest average claim amount for the largest number of procedures (2 versus 1 and 1).
    - Which **three** providers had the largest *claim difference* for the largest number of procedures, where the claim difference is the difference between the average amount claimed by a provider for a procedure and the average amount reimbursed for that provider and procedure.
      To clarify, consider the following example. If out of two providers, if Provider A has the largest claim difference for procedures 1 and 2, and Provider B has the largest claim difference for procedure 3, then Provider A has the largest claim difference for the larger number of procedures (2 versus 1).
2. Some providers and regions are likely to be different in more subtle ways. Based on the data provided, which **three** providers are **least like** the others? Briefly explain what seems to be different about these providers. Which **three** regions are **least like** the others? Briefly explain what seems to be different about these regions.
3. Medicare staff have identified a number of records in the individual patient claims data that look unusual – they could have errors or fraudulent claims or simply be unique patient contexts that are worth review. Using the list of unusual records hand-selected by the staff as a guide, identify **10,000**

additional patient recordsthat seem most likely to also need review. Briefly describe some common features in these patients.

Deliverables

Each challenge submission must consist of a compressed tarball or zip file with the following contents:

Part 1 Solution

- The ICD–9 codes of the three procedures with the highest relative variance, one per line, in a **text file** called part1a.csv
- The IDs of the three providers with the highest claims on the most procedures, one per line, in a **text file** called part1b.csv
- The three locations (e.g., cities) and states with the highest claims on the most procedures, one per line as "*location,state*" (where *state* is the two-letter abbreviation, e.g., *CA* for *California*), in a **text file** called part1c.csv
- The IDs of the three providers with the largest difference on the most procedures, one per line, in a **text file** called part1d.csv

Part 2 Solution

- The IDs of the three providers that are least like the others, one per line, in a **text file** called part2a.csv
- The three cities and states that are least like the others, one per line as "*city,state*" (where *state* is the two-letter abbreviation, e.g. *CA* for *California*), in a **text file**called part2b.csv

Part 3 Solution

- The IDs of the 10,000 patients whose records most need review, one per line, in a **text file** called part3.csv

Solution Abstract

Brief write-up in **PDF format** that addresses the following points:

- Part 1
    - Explain your methodology including approach, software and algorithms used, any testing and validation techniques applied, and total time spent.
- Part 2
    - Explain your methodology including approach, software and algorithms used, testing and validation techniques applied, model selection criteria, and total time spent.
    - Explain how/why the three providers are different.
    - Explain how/why the three regions are different.
- Part 3

- Explain your methodology including approach, software and algorithms used, testing and validation techniques applied, model selection criteria, and total time spent.
- Explain why the 10,000 patients are worth a manual review.

Please include in your solution abstract any information that can be used to understand the logic behind your approach and all steps taken, including data preparation, modeling, validation, analysis, visualization, etc. The solution abstract should typically be 3 to 5 pages and **no more than 6 pages**.

Complete Source Code

- Tarball or zip file of **all source code** used to complete the challenge, including programs, scripts, and other artifacts.

Challenge Submissions

You may submit your solutions at any point prior to the submission deadline. You may submit as often as you like, but only the final submission will be scored. No submissions will be scored until the submission deadline has passed.

Each submission should be uploaded to this drop box using the HTML submission form. You will receive a confirmation within one business day of submission.

Bear in mind that uploads times will depend on the speed of your connection and the size of your submission. You are therefore strongly encouraged to submit your solution at least two hours before the submission deadline.

Submission Scoring

Submission Deadline: June 30, 2014, 23:59:59 US Pacific Daylight Savings time (UTC-08:00)

Submissions will be scored as follows. Each problem part will be scored independently. The score for each part will be a composite of the percentage correct for all submitted solutions for that part and the score assigned to the corresponding section of the solution abstract. The scores for the three parts will be weighted and combined into a final composite score. The third part will have the greatest weight, and the first part will have the least weight.

The percentage correct for each part will be scored against a golden master of known correct answers. Note that some questions may have more than one correct answer.

The solution abstract will be scored according to objective criteria about your approach and general mastery of the tools and techniques. Writing quality and formatting will not contribute to the score, except in cases where the writing is so poor as to impact understanding.

Execution Environment

The recommended execution environment is an Apache Hadoop cluster (such as CDH4 or CDH5) with at least four nodes. If you do not have access to local cluster resources, the Amazon EC2 cloud may be a good alternative. Instructions for installing a Hadoop cluster in EC2 using Cloudera Manager are available

in the product documentation. For access to free cloud computing resources, try the Verizon Cloud beta program, which is available until the end of May.

If you have sufficient processing power and free storage space, it may also be possible to work only on your local system, either using local tools or a Hadoop cluster in pseudo-distributed mode. To avoid setting up a local pseudo-distributed cluster, you can download and use the Cloudera Quickstart VM. Please note that some parts of this challenge may require significant computing resources, so use of a local environment or a virtual machine may limit your effectiveness on the challenge.

Useful Links

- Data downloads:
    - http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient.html
    - http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Outpatient.html
    - http://cloudera.box.com/s/7h2sx81w73mm1ciswt5f
- Challenge Forum
- Download Cloudera Manager and Cloudera's Distribution including Apache Hadoop
- Cloudera Quickstart VM
- Apache Crunch
- Cloudera ML
- Data-Intensive Text Processing with MapReduce
- Mining of Massive Datasets
- http://en.wikipedia.org/wiki/Diagnosis-related_group
- http://en.wikipedia.org/wiki/Ambulatory_Payment_Classification
- http://en.wikipedia.org/wiki/Delimiter#ASCII_delimited_text
- http://www.bcbsnc.com/assets/providers/public/pdfs/837Professional_5010_v2.9.pdf

Hints and Tips

- Get started early! This challenge will take a significant amount of time to complete. Especially if you're doing it in your spare time, get started as early as you can.
- You can submit as many times as you like, but only your last submission will be scored.
- If you're having trouble understanding any part of the challenge, try searching the challenge forum or posting a question there.

Rules

Individual Contributions Only
You must participate in this challenge only on an individual basis; teams are not permitted.

Sharing
Any sharing of code or solutions or collaboration with another person or entity is strictly forbidden.

Tools
You may use any tools or sofware you desire to complete the challenge.