REMINDER: DO NOT DOWNLOAD OR COPY ANY DATA FROM THE IBM SERVER.

6.S897/HST.S56 Problem Set 2
Due: **Tuesday March 5 by 11:59pm** through stellar

Instructions

Log in to your account on the IBM server. Contact the TAs on piazza if you have not received login information for your account yet. You will have your own virtual server, though each server is single core, so do not run many jobs at once.[1]

You will find a *ps2materials* directory in your home folder on the IBM server. This contains starter code. You do not need to submit your code on stellar.

In the body of this exercise, blue text describes what specifically you should be submitting and orange text describes fourth-wall-breaking commentary from the TAs. If you have any questions about what the wording of the questions mean, ask us on piazza!

2.0: Setting the Scene

Dr. Willow Rosenberg, the new Director at the Consortium for Disorder Control (CDC), cares strongly about using data to improve the health of Americans. She heard that you excelled at HST.956 and invited you to visit the CDC to talk with some of the staff to suggest data-driven solutions. She asked that for your final assessment, you should a standalone report of your analyses with relevant plots, interpretation of findings, and recommendations named ${mit_user}.pdf (e.g. *wboag.pdf*). Do not copy your code from the IBM server or submit it.

In your write-up, please make your answers as easy to identify as possible (e.g. highlight with colors, label questions with numbers, etc). The faster we can identify your answers, the faster we can grade 70 submissions! :)

Part 1: We Have All This Data! What Do We Do With It?

You first meeting with Dr. Rupert Giles, the Deputy Director for Public Health Science and Surveillance. He believes that in order to prevent diabetes, the CDC needs to do a better job catching it early. A predictive model could allow for better early detection and intervention. Through a new partnership with the Institute of Bagels and also Medicine (IBaM), they have new access to a large dataset[2] of commercial claims, with information covering both inpatient and

[1] Servers have 8G of RAM. You likely will not be able to run all three notebooks at once, depending on how efficient your solution is. You can monitor your MEM usage with “htop”.

[2] Note that for this pset, we downsampled the number of non-diabetic patients from 800k to ~40k. This means the starting cohort is not representative of the true population.

outpatient services as well as prescription drugs. He knows the data is useful, though he isn't sure what the best way to operationalize a program like this would be.

The two of you visit the office of Dr. Tara Maclay, the Deputy Director for Public Health Service and Implementation Science. Dr. Maclay has experience successfully deploying projects, so you excitedly listen to her advice. She thinks that you should develop a predictive model which uses data collected about patients from 2011 to 2012 to predict whether a given patient will develop diabetes in the year 2014. Any patients who develop diabetes before 2014 or after 2014 should be excluded from the prediction cohort.

Go to notebook "Q1. Cohort Creation.ipynb" and follow the directions there. Make a table in your write-up indicating whether each of the following patients will be included in the prediction task's cohort, and why (in one sentence). For instance (first row provided as example):

ENROLID	Included? (Y/N)	Why?
107359602	No	Diabetes onset during the collection period (on April 2011).
176051105		
177907802		
201493606		
33533505		
381209601		

Create your cohort based on Dr. Maclay's advice. Describe your cohort in the following ways:

1. How many patients are in the final cohort? How many were excluded?
2. In the final cohort, what fraction of patients are positive examples?
3. Fill this table which mirrors some of the categories of Razavian et. al's Table 1.

characteristic	Total Population	Population with diabetes
Number of patients		
% Female		
% patients with hypertension (ICD9 of 401.xx)		

We do not want this first part to be a time sink or blocking point for this assignment, so for Parts 2 and 3, we provide you with the cohort. We want you to get experience trying to formulate this machine learning problem, but please don't spend more than 2.5 hours on it. If you do not finish in time, note that and report how far you got. If you are struggling with setting this problem up, we encourage you to make use of piazza and office hours.

Part 2: Building a Predictive Model

For this question, use the cohort loaded by the notebook "Q2. Logistic Regression.ipynb" regardless of whether you completed Q1. This will standardize grading by making sure everyone has the same train/test splits.

After following Dr. Maclay's advice, you reconvene with Dr. Giles. He asks you for a simple baseline model: L1-regularized logistic regression on demographics (demo) as well as the presence/absence of each ICD/NDC code. You do not need to make multiple "windows" of times, simply use an indicator of whether each event occurred during the collection period. An "event" should contain everything printed in one row of chart_review, except for the timestamp.[3] All features should be categorical. We provide the DivctVectorizer and model code so you only need to write the feature extraction. The model is very fast to train (~10 seconds to fit the LogisticRegression model).

In your write up:

- For C=0.1
    - Report the AUC of that model.
    - Report the top 5 features most associated with developing diabetes.
    - Pick 3 of the 5 features and explain why it is plausible the model identified them.
- Complete this table with the AUC and number of nonzero weights in the learned model.
    - Reflect on what these feature and regularization experiments indicate about the predictive power and redundancy of claims data.

features	LogisticRegression C	AUC	Number of nonzero weights
NDC and demo	0.02		
NDC and demo	0.1		
ICD and NDC and demo	0.01		
ICD and NDC and demo	0.02		
ICD and NDC and demo	0.1		
ICD and demo	0.01		
ICD and demo	0.1		

[3] An example event would be: ('prescription_drugs', 'ndc', '93104801').

Only demographics	0.02		
Only demographics	0.1		

Part 3: What Could Go Wrong?

Dr. Giles is very excited that your tool seems to be working. He sends you to Dr. Maclay's office to discuss the best way to deploy this predictive model as soon as possible. However, once you arrive to Dr. Maclay's office, she seems concerned about how quickly Dr. Giles wants to be moving forward with this project. She insists that the model has not been thoroughly tested and is not ready to be deployed. She knows that Dr. Giles can be stubborn and will want convincing evidence that there's much difference in the data over time. She gives you access to some data from 2015 and recommends that you find an example of "dataset shift." See the notebook "Q3.Dataset Shift.ipynb". She assures you that once you write that up, her team can take over from there and work further with Dr. Giles about testing and deployment.
In your report:

- *Plot a histogram (binned by month) of the number of patients that have an ICD9 code of 250.xx during that month. There should be 24 bins spanning from January 2014 to December 2015.*
- *Describe what this plot says about dataset shift.*

You gather your things after a productive day of meetings and say your goodbyes. Dr. Rosenberg thanks you for your time and asks that you send her your report within one week.