**Instructions**

You have 4 hours to complete the exam.

You may use Excel or any other program(s) of your choice (R, Stata, SPSS, SAS, etc.) to assist you. You may NOT ask anyone else for help with this exam. You may NOT include any content that is not your own. Also, do NOT include any content from previous work, classes, etc. All answers must be original.

For this test you will answer several questions that require written response. Create a word document named yourname\_datascience1\_written.docx and paste responses into this document with the question text above the response. This document will be referred to as the written portion for the remainder of this exam.

Suggested timing for each question is just a guide – you may spend more time with some questions and less time with others, provided that you return the completed exam within four hours after the exam is emailed to you.

Question 1: 150 minutes

Question 2: 15 minutes

Question 3: 45 minutes

Question 4: 30 minutes

**Question 1** [150 minutes suggested]:

You have been tasked with producing an individual level prediction of a person’s probability to have health care. In the attached dataset you will find several hundred thousand people with known health coverage information (hicov) and nearly 200,000 people with unknown health insurance information. Use the attached data set to build individual level probability predictions for the people for whom we have unknown health insurance coverage information. You can find a data dictionary for the included data [here](http://www.census.gov/acs/www/Downloads/data_documentation/pums/DataDict/PUMS_Data_Dictionary_2009-2011.pdf):

* Explain your choice in methodology in the written portion;
* Explain how you used your chosen method to solve the problem in the written portion;
* Paste all of your code used in solving this problem into a text file named yourname\_datascience1\_code.txt and send back;
* Paste your final model output into a text file named yourname\_datascience1\_output.txt
* If you want to send additional output that does not translate into a text file, send the file using the following nomenclature: yourname\_datascience1\_output\_typeofoutput;
* Finally please send all records that you’ve scored as a CSV file named yourname\_datascience1\_scores.csv:
  + Two numeric columns only
  + Please include headers: aa\_primarykey, probability\_score

**Question 2** [15 minutes suggested]:

A new non-profit health insurance cooperative in Oregon has purchased the model you built. The model is called the Uninsured score, ranges from 0-100, and represents the probability this individual **does** **not** have insurance.

You have been in frequent contact with an analyst at the cooperative, helping him run analysis on the scores. One day he asks you curiously:

*“According to your uninsured model, my friend has an uninsured score of 15 (out of 100), but I know for a fact he does not have insurance and hasn’t for a long time. Why has your model assigned him a score of only 15?”*

Without knowing this friend’s individual information, how do you explain this apparent discrepancy to the analyst? Assume the analyst is a smart, educated person with little or no background in statistics.

Then suggest a different way that the analyst could confirm the accuracy of the predictive model created by your colleagues.

Include your answers as a part of the written component of this test.

**Question 3** [45 minutes suggested]:

It is summer of 2013, and you are a Civis Analytics representative attending a conference of non-profits, charities, doctors, and health insurers. The theme of the conference is how the Affordable Care Act is changing the landscape of health care in the United States and how each sector of the health care industry can prepare for it.

Being the proud employee you are, you chose to wear a Civis Analytics T-shirt. Preston Tucker, a CEO attending the conference, recognizes the Civis logo and comes over to you to ask:

*“We are Colorado Health Council, the largest health insurance company in the state of Colorado, but we currently only work with employers. We are planning to offer individual insurance on the Colorado insurance exchange starting in 2014, but we don’t know where to start. What are the things we should know and be considering as we make this move to a new market?”*

Eyeing a potential partner, you decide to skip the coffee and networking session and give him an answer with supporting evidence by the time everyone returns in 45 minutes.

Use the attached materials (“Question 3 Background Materials.pdf”) to send a response as an email addressed to this CEO with 2-4 short bullets, backed up by supporting evidence. **You do not need to consult any other references.** We recommend you begin by reading the introductory document “The ACA and Insurance Exchanges,” which contains a glossary of key terms and a basic overview of how the Affordable Care Act will affect the insurance market.

Please cite your references. Include your answers as a part of the written component of this test.

**Question 4** [30 minutes suggested]:

Thanks to your excellent handling of Question 3, the Colorado CEO introduces you to a friend of his who works at a large charitable foundation. They both agree that finding the people who are uninsured is one of the biggest remaining challenges to ensure that the Affordable Care Act is a success. Knowing in advance who is likely to be uninsured means that non-profits could better target their services, the insurance companies could market specifically to those people, and governments could pick smarter places to place public service advertising and outreach.

You decide Civis should help find the uninsured in the United States. The project is to develop an individual-level model that predicts the likelihood the individual does not have health insurance.

You have access to the following:

* A national data file with a record for 260 million individuals in the country – including their name, address, phone number, demographic information (age, race, ethnicity), and a few estimated consumer variables (income, car ownership, home ownership). It has no health data. This is already loaded into an accessible database.
* A Northeast regional health conglomerate’s data file, which includes name, address, phone number, and medical history, including health insurance status, for 55 million people who have used health services in Northeast region clinics or hospitals in the past 25 years. We do not know the quality, timeliness or completeness of this dataset. This is also loaded into the same database.

Before bringing analysts or data scientists onto the project, you are asked to assess the contents of the data provided to you. What would you do in your 1st week to organize the data, assess its quality and usefulness and communicate the results to others?

Include your answers as a part of the written component of this test.