**Homework 1**

**Due: 5pm April 4, 2016**

You are being provided with data in a CSV file that contains student records. Each row is a record about a student containing their name, demographic information, grades, test score, attendance, and whether they graduated or not.

Problem A

1. The first task is to load the file and generate summary statistics for each field as well as probability distributions or histograms. The summary statistics should include mean, median, mode, standard deviation, as well as the number of missing values for each field.
2. You will notice that a lot of students are missing gender values . Your task is to infer the gender of the student based on their name. Please use the API at [www.genderize.io](http://www.genderize.io) to infer the gender of each student and generate a new data file.
3. You will also notice that some of the other attributes are missing. Your task is to fill in the missing values for Age, GPA, and Days\_missed using the following approaches:
4. Fill in missing values with the mean of the values for that attribute
5. Fill in missing values with a class-conditional mean (where the class is whether they graduated or not).
6. Is there a better, more appropriate method for filling in the missing values? If yes, describe and implement it.

You should create 2 new files with the missing values filled, one for each approach A, B, and C and submit those along with your code.

Please submit the Python code for each of these tasks as well as the new data files for this assignment.

Problem B

A larger data set than the one in the previous problem was used to build a *logistic regression* model that predicts the probability an individual student will graduate.

Below are coefficients from this model.  The definitions of the variables are below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **coefficient** | **standard error** | **z score** |
| **Male** | 1.45 | 0.09 | 15.81 |
| **Female** | -2.11 | 0.1 | -21.95 |
| **Ln\_Family\_Income** | -0.109 | 0.041 | -2.63 |
| **Age** | -0.013 | 0.0096 | -1.4 |
| **Age\_Sq** | 0.0001 | 0.00009 | 1.52 |
| **Census\_College** | 1.77 | 0.33 | 5.37 |
| **AfAm** | 2.07 | 0.399 | 5.18 |
| **AfAm\_Male** | -0.872 | 0.437 | -2.01 |
| **Constant** | 1.2 | 0.484 | 2.48 |

* Male—Coded as 1 if the student is M, 0 if he/she is not
* Female—Coded as 1 if the student is Female, 0 if he/she is not
* Ln\_Family\_Income—The natural logarithm of the student’s Family income (in dollars)
* Age—The student’s age (in years)
* Age\_Sq—The student 's age (in years) squared
* Census\_College—The percentage of residents in the student ‘s neighborhood who have a college degree (scaled from 0 to 100)
* AfAm—Coded as 1 if the student is African American, 0 if he/she is not
* AfAm\_Male—Coded as 1 if the student is both African American and Male, 0 if he/she is not
* Constant—The constant term

1. Consider 4 students, Adam, Bob, Chris and David. Adam and Chris share identical characteristics except for their family incomes.  Bob and David also share identical characteristics (with each other, not necessarily Adam and Chris), except for their incomes.

|  |  |  |
| --- | --- | --- |
| **Name** | **Family Income** | **Modeled Probability of Graduation** |
| Adam | $50,000 | 50% |
| Bob | $200,000 | 50% |
| Chris | $40,000 | ? |
| David | $190,000 | ? |

Based on the coefficients above, who would you think has a higher probability of graduating?

* Chris
* David
* They have the same probability
* Cannot tell based on the information provided

What is your reasoning?  (you need not calculate an exact probability to answer this question. Just explain your reasoning in general terms.)

1. The coefficient for AfAm\_Male is negative. How do you interpret this? Does this mean that African-American Males are more likely to not graduate than African-American Females? What about relative to non African American males?
2. How do we interpret the difference in graduation probability between students of different ages? How do the variables in the model estimate such probability?
3. Are there any variables in this model that you would choose to drop? Why or why not? Would you need more information in order to make this decision?

Instruction of how to submit your homework is found at [http://people.cs.uchicago.edu/~larsson/capp30254-spr-15/](http://people.cs.uchicago.edu/%7Elarsson/capp30254-spr-15/). The instructions assume familiarity with the submission system used in CMSC 12100 and 12200. If you did not take these classes, please come to office hours for help getting your assignment submitted.