**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;

Suppose the value of 1 for hicov indicates an insured person and the value of 2 indicates an uninsured person. Just like any other classification problems, we can use different classifiers to train the model, then verify the validity of the classifier using cross validation. Those classifiers include Naïve Bayes, Support Vector Machine, Random Forest, Decision Trees and K nearest neighbors. However, since we are predicting a person’s probability to have healthcare, the logistic regression makes much more sense to the other classifiers because it can give us a precise estimation of the probability, while other classifiers will just give predictions on whether the person is likely to be insured or not, but not the exact likelihood of that.

I used 5-fold cross validation. We first divide the dataset into a useful set and a prediction set, the useful set contains all rows which have hicov values. We will use the useful set to apply cross-validation which will give us the model with the best test error, then predict the probability on the prediction set. I used 0-1 loss function and will select the model with the lowest error rate. Applying the logistic regression model,