Prediction of influenza vaccination status for Michigan residents via classification methods

Stewarjr

ABSTRACT

Objective Influenza vaccinations have become more widely available via clinics, pharmacies, and even college health centers or employers but the number of Americans that have not been vaccinated is still about half the population [1]. Using data from the Centers for Disease Control and Prevention (CDC), models were constructed to predict if a Michigan resident would get one of two common influenza vaccinations: a flu shot or the nasal spray.

Methods Using the BRFSS data for the years 2011-2013, classification models were constructed on the training set which consisted of the years 2011 and 2012 while the 2013 data set was held out to be treated as a test set. The classification models implemented to predict the dichotomous variable of whether or not a Michigan resident received a flu vaccination were logistic regression, linear discriminant analysis, k-nearest neighbors, and a decision tree.

Results Of the four classification methods implemented, logistic regression and linear discriminant analysis had the highest prediction accuracies of 65.03% and 65.06% respectively. With the use of random forests for variable selection, it appeared that three variables were of high importance: the status of an individual had or had not taken the pneumonia vaccination, the annual household income, and the employment status of the individual. When a logistic regression was run on just these three variables, the prediction accuracy was 64.68%.

Conclusion If one model were to be selected for the prediction of a Michigan resident's flu vaccination status, it would be logistic regression. One reason why this would be the case is because the majority of the predictors used were categorical in nature – with a few quantitative discrete variables – which works fine with the assumptions of a logistic regression. Unfortunately, the multivariate Gaussian distribution is an assumption of a linear discriminant analysis which means the predictors should be continuous. Also, k-nearest neighbors suffers from the lack of quantitative variables as well since the default distance measure is the Euclidean distance. Lastly, the decision tree was based on one split because the predictor for pneumonia vaccination status was a significant classifier for flu vaccination status. Therefore, the logistic regression had the most rigid foundation for the prediction of flu vaccination status due to its less rigorous model assumptions.

BACKGROUND

Over the past 31 influenza seasons, the number of deaths ranged from a minimum of 3,000 people to about a maximum of 49,000 people. Of these deaths, 90% of the people were an age of 65 years or older [1]. It is recommended to get the flu vaccination in early October since the vaccination takes

about two weeks before it becomes effective against the virus and the flu season can start as early as October and as late as May. With more people getting the vaccination, there would be a lowered risk of developing influenza because of a phenomenon called herd immunity that would limit the transmission and dispersion of the influenza virus.

The CDC recommends that every aged 6 months or older should get the flu vaccination every season; this vaccine recommendation has been established since the Advisory Committee on Immunization Practices (ACIP) voted for universal in the United States in February 2014 to increase the available of both the flu shot and nasal spray [1].

METHODS

The Behavioral Risk Factor Surveillance System (BRFSS) datasets for the years 2011, 2012, and 2013 were obtained from the CDC website. From the 2013 dataset alone, there was over 500,000 observations with 330 variables to consider. Thus, a subset on Michigan residents was established to reduce the size of the data to allow all of the data to be read in memory for the statistical program R. 2013 was the latest dataset found on the CDC website; therefore, this year will be the test set in which models will applied to predict the influenza vaccination status of a Michigan resident. The years 2011 and 2012 were combined to create the training set; the models were made from this training set.

To get the data in R, the three SAS XPORT files were downloaded to my R directory and read in as data frames via the *foreign* library. One feature that was immediately noticed when the dataset was read in memory was that each year had a different number of variables. This indicated that the CDC was not consistent with the survey over the recent years. When reading the survey methodology, it becomes clear that the survey is slowly transitioning from being conducted through landline phones to mobile phones to accommodate for the decline in landline phones [2]. During the transition of the survey methods for including cell phone users, the number of variables decreased from 434 in 2011 to 330 in 2013. Thus, my method was to only consider variables that appear in every dataset. However, more difficulties appeared when validating the data has been read correctly via summary statistics. For instance, new inconsistencies found within the CDC datasets were slight variations on variable names and factor level labels. In the 2013 dataset, the variable of interest – influenza vaccination status – is called FLUSHOT6 in the 2013 dataset while it was called FLUSHOT5 in the 2011 and 2012 datasets. Most of the variable name variations were found with ease via string matching functions. An example of a factor level label inconsistency was for the DISPCODE variable. This variable indicated whether or not the person completed the entire survey. In the 2013 dataset, 1100 indicated that person completed the interview and 1200 indicated that a partial interview was recorded. The other two datasets simply had 110 and 120 for completed and partial interviews respectively. The remedy to this difference was to just use the same labels as the latest year.

Another peculiarity was found when assessing the form of missing data (*i.e.* is the data missing completely at random, missing at random, or not missing at random). For instance, the variable MSCODE describes

an individual's vicinity to a metropolitan population density. However, a missing value for this variable did not indicate that the individual failed to address the question, it indicated that the individual was in a United States territory such as Guam, Puerto Rico, and the Virgin Islands. Also, every observations that had a missing value for MSCODE had a missing value for

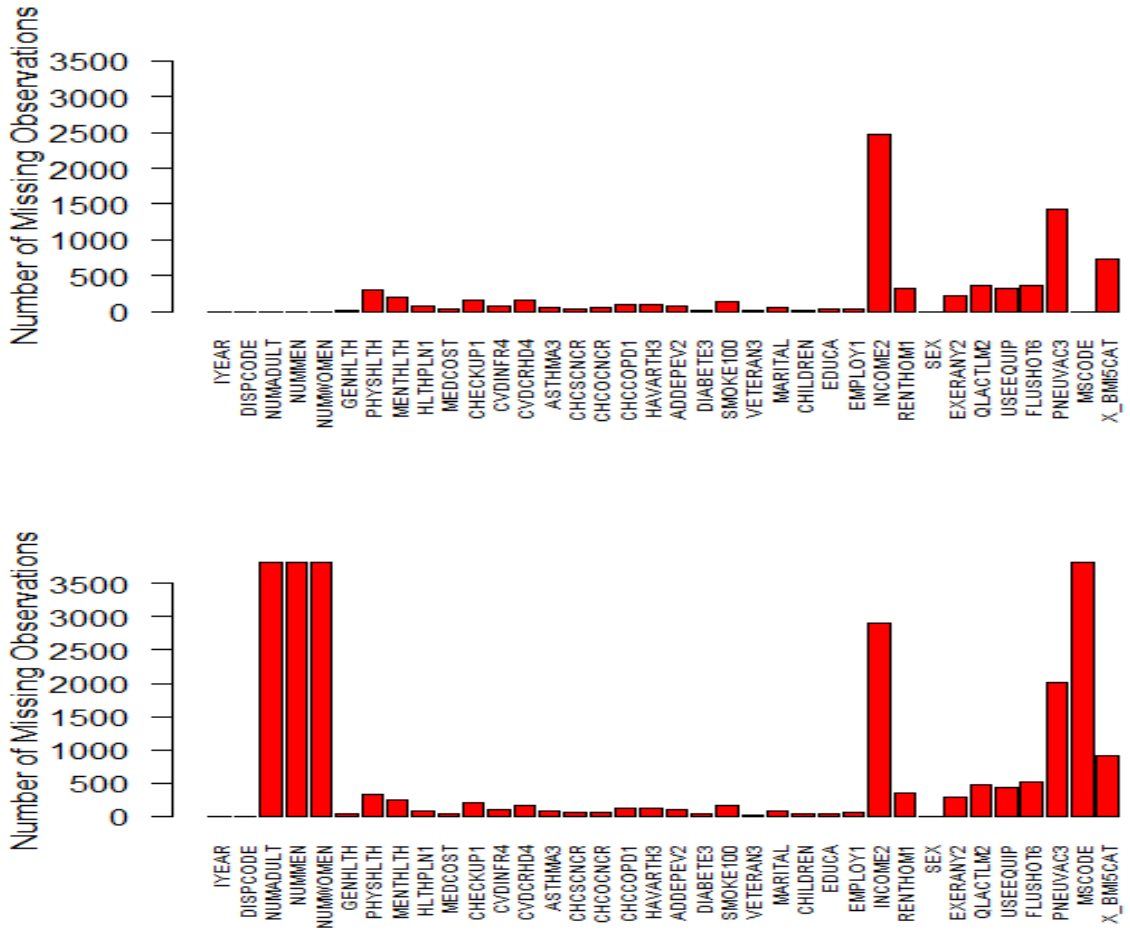

Figure 1. The number of missing observations in the training set (years 2011 and 2012) were constant for the variables NUMADULT, NUMMEN, NUMWOMEN, and MSCODE. From the BRFSS codebook, a missing value (NA) for MSCODE meant that the individual was from Guam, Puerto Rico, or the Virgin Islands. These observations were removed for not being Michigan residents.

NUMADULT, NUMMEN, and NUMWOMEN. Each of these variables represent the number of adults, men, or women in a given household respectively. Therefore, these observations were removed from the data sets as indicated in Figure 1 since it is questionable whether or not the individual is a Michigan resident since the individuals living in these U.S. territories did not have these variables recorded.

Next, variables missing enough observations were removed because eventual imputations of this missing data may have been extraneous when the variable may not be a significant predictor of flu vaccination status in the first place. The threshold used here was if half of the observations were missing, then the predictor was removed. This brought the number of predictors down to 163 when using half of the observations missing as a threshold. The majority of the variables were removed because their use was for the identification of the individual who took the BRFSS survey such as phone information (which was given a unique ID to prevent privacy concerns) and survey stratification and weighting which left about 40 variables.

Before one can impute missing values for the predictors that are left in the data set, an analysis of the missing data has be implemented because how your data is missing determines what imputation method will be used. Missing data can typically be classified as completely missing at random (CMAR), missing at random (MAR), or not missing at random (NMAR). If the data are CMAR or MAR then the imputation method is simpler because there would be a lack of an association between missing values of predictors that can be treated as if the missing value is a random variable instead of a conditional variable that would increase the difficulty of modeling the missing data. An example of MAR would be many countries taking a survey about education in which some countries would not be able answer what their GPA was because this metric does not make sense to them since it was not used; therefore, the respondent would likely leave this question blank. For NMAR, there would have to be specific demographic characteristics that would deter the individual from wanting to responds to the question; an example would be asking the number of hours for exercise per day to an obese person since they may not admit that they are lacking in exercise so they would leave the question blank. To observe how your data is missing, a matrixplot can be constructed for this assessment.

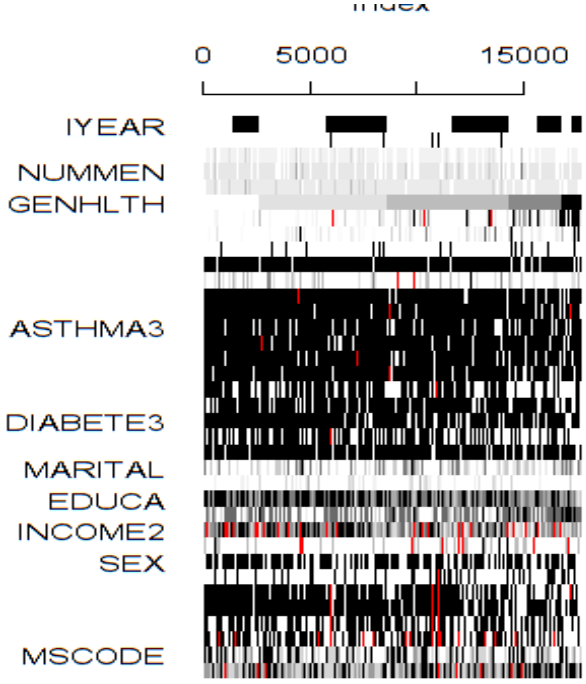

Figure 2. Sorting the data by GENHLTH shows that the missing values denoted as red appear to have a random scatter to suggest the data is MAR.

In Figure 2, the data were sorted by the general health status of an individual. This variable was chosen because of the multiple classes that it could take on and its possible relationship to other variables to potentially see if those with poorer health would be less inclined to answer questions that relate to their health stature. It appears that even after sorting by general health, no pattern is evident for the missing data denoted in red; therefore, the data is likely to be MAR.

Since the data was determined to be missing at random, the multiple imputation used was multivariate imputation by chained equations. This was accomplished by using the R package *mice* to generate a new data set that has imputed values for any observation that was missing. The general method of *mice* is to iterate through each column/predictor and creating a model that conditioned on every other column. For example, if the column that is currently being imputed is a 3-level categorical variable, then a polytomous logistic regression is run with every other column as a predictor. This continues until each predictor has been evaluated five times to create the best, reasonable estimate for the missing data. The final output data set consisted of 33 predictors and 17,693 observations in the training set and 8,402 observations in the test set.

Once the training and test sets were imputed, the classification models were applied. The models considered for the prediction of flu vaccination status were logistic regression, linear discriminant analysis, k-nearest neighbors, and a decision tree. All 33 predictors were used in the four different models and then a random forest model was created as well to determine which of the 33 predictors had the greatest contribution for the prediction of flu vaccination status. Then, for each model a confusion matrix was generated to calculate the prediction accuracy (or the complement of the misclassification rate).

ANALYSIS/RESULTS

For each classification model, a confusion matrix was constructed to determine the number of correct predictions. A confusion matrix is essentially a contingency table on these two categorical variables: the actual/test influenza vaccination status and the predicted influenza vaccination status. The prediction accuracy is determined from a contingency table because our response variable is a binary categorical variable with levels “yes” and “no”. For the logistic regression, the percentage of correct predictions for influenza vaccination status was (2340 + 3124)/8402 = 65.03%; this model is somewhat better than random guessing.

A linear discriminant analysis was utilized, but the assumptions of the model were likely not met. That is, the predictors do not resemble a multivariate Gaussian distribution since the majority of the predictors are categorical. However, the linear discriminant analysis was used for prediction anyway. The confusion matrix yielded a prediction accuracy of 65.06% which is also somewhat better than random guessing and comparable to the multiple logistic regression method. Since the model assumptions were not met, it is not meaningful to interpret the generated parameters and the logistic regression may be the preferable model to conclude with thus far.

K-nearest neighbors used a value of k=5 and this yielded a prediction accuracy of 58.51%. This model should be considered lightly because its use of the Euclidean distance measure on this categorical data limits the inference that can be drawn from this model.

The last method used was a decision tree and this yielded a prediction accuracy of 64.66%. However, the tree that was constructed only had one split because the predictor for the pneumonia vaccination status of an individual contributed greatly to the prediction of flu vaccination status.

Flu vaccination status prediction (33 predictors)				
Classification Model	Logistic Regression	Linear Discriminant Analysis	k Nearest Neighbors (k = 5)	Decision Tree
Accuracy	65.03%	65.06%	58.51%	64.66%

From the random forest, it was found that the variables PNEUVAC3, INCOME2, and EMPLOY1 were deemed "important" by having the greatest mean decrease Gini indexes (PNEUVAC3 being higher than INCOME2 and EMPLOY1). These variables indicate the pneumonia vaccination status of an individual, the annual household income, and the employment status respectively. Since the logistic regression most suitable model for this data, this model was run again with just these three predictors. The logistic regression with these three predictors yielded a prediction accuracy of 64.68% which is just about as good as the models that had all 33 predictors.

DISCUSSION

Essentially, with the data given, the models were only somewhat better than random guessing – of which, logistic regression objectively did the best. From the random forest variable selection by importance, the PNEUVAC3, INCOME2, and EMPLOY1 variables had the greatest impact on determining the flu vaccination status of a Michigan resident. This suggests that future studies could examine why these variables are associated with the desired outcome. Some speculations of how employment is involved include individuals not wanting to get the flu vaccination because that is the least of their worries in the current situation despite the increased availability of vaccinations via pharmacies, clinics, etc.

References

(1) Seasonal Influenza. 2014; Available at: http://www.cdc.gov/flu/protect/keyfacts.htm Accessed November 17, 2014.

(2) The BRFSS Data User Guide. 2014; Available at: http://www.cdc.gov/brfss/data_documentation/PDF/UserguideJune2013.pdf. Accessed November 21, 2014.