STAT 652 Final Project

Arin Ghosh

12/1/2018

1. Introduction

This is an analysis report of the data provided by the Canadian Community Health Survey (CCHS) – Healthy Aging module. The project is divided into 2 parts corresponding to two separate datasets provided to us. The first dataset has 20000 rows and 9 columns. In this dataset problem our task is to predict cognitive health index called HUIDCOG using 8 other health-utility-index (HUI) variables. In the second part of the project, we are tasked with building a regression model that predicts a real number called HUIDHSI, which is another measure of the HUI that provides a description of an individual's overall functional health. The second dataset is much bigger in size compared to the first dataset, having 590 variables with 10000 rows. A part of the dataset is held out for validation purpose which was released to the students at a later date.

Part 1: Predicting HUIDCOG (Classification Analysis)

1 Data

1.1 Data Loading

First step is to load the appropriate dataset into the R Studio environment. The dataset can be found on the project github repository. Once downloaded in to the working directory of the R Studio, we can load the data using read.csv() command.

```
hui <- read.csv("hui.csv")
summary(hui)
```

```
##            DHHGAGE       DHH_SEX                  HUIDCOG
##  55 TO 59 YEARS:3085   FEMALE:11385   COG. ATT. LEVE 1:13949
##  60 TO 64 YEARS:2982   MALE  : 8615   COG. ATT. LEVE 2:  496
##  85 AND OLDER  :2602                  COG. ATT. LEVE 3: 3764
##  65 TO 69 YEARS:2595                  COG. ATT. LEVE 4: 1268
##  70 TO 74 YEARS:1958                  COG. ATT. LEVE 5:  429
##  75 TO 79 YEARS:1928                  COG. ATT. LEVE 6:   71
##  (Other)       :4850                  NOT STATED      :   23
##             HUIGDEX                  HUIDEMO                 HUIGHER
##  LIM. HANDS/F   :  252   EMOT. ATT. LEV.1:14912   NO PROBLEMS    :17335
##  NOT STATED     :   10   EMOT. ATT. LEV.2: 4067   NOT STATED     :  296
##  USE OF HANDS/F.:19738   EMOT. ATT. LEV.3:  749   PROB./CORR.    : 1579
##                          EMOT. ATT. LEV.4:  183   PROB./NOT CORR.:  790
##                          EMOT. ATT. LEV.5:   39
##                          NOT STATED      :   50
##
##             HUIGMOB                  HUIGSPE                  HUIGVIS
##  NEED MECH. SUPP: 1580   NO PROBLEMS     :19837   NO PROBLEMS     : 4210
##  NO AID REQUIRED:  322   NOT STATED      :   11   NOT STATED      :  142
##  NO PROBLEMS    :17496   PARTIAL/NOT UND.:  152   VISUAL P. UNCOR.:  658
##  NOT STATED     :   16                            VISUAL PROB. COR:14990
```

```
##  REQUIRES HELP   :  586
##
##
```

From the summary of the hui dataset, we can see that there are a total of 9 columns that exists. Further, we see from the summary that there are no NA or missing values, although there are several entries with value 'NOT STATED'. Also it is noteworthy to mention that none of the variables are continuous variables and the value which we are suppose to predict is a multivariate i.e. HUIDCOG can take one of the possible 7 values for each set of dependent variables. Although the dataset appear to be clean and variables are grouped, we will still need to do some analysis and take a look if we can reduce possible number of outcomes for each column without loosing a lot of quality in the dataset where it makes sense.

1.2 Exploratory Data Analysis & Data Grouping

Missing Data:

To start with our data analysis, we can notice that for some of the responses, we have an inconclusive response i.e. `hui$HUIDCOG== 'NOT STATED'`. We should first remove these records, since this leads to an observation where we don't know what is the outcome.

```
# remove NOT STATED from HUIDCOG
hui <- hui[hui$HUIDCOG!= 'NOT STATED',]
hui$HUIDCOG <- factor(hui$HUIDCOG)
```

Target Variabe:

The given dataset is a fully categorical mulltivariate dataset meaning there are no columns with real numbers. This is something that is not extinsively discussed in our class, although we've been tought how to deal with categorical data in general. Since multivariate analysis is not extensively covered in the coursework, I am going to reduce the possible outcome of our dependent variable HUIDCOG from 6 to 2. To come up with a meaningful division, we must refer to the original documentation provided by the instructor. If we take a look into page 53 of CCHS_HA_Derived_variables.pdf then we can find out how these 6 different classes of HUIDCOG came into existence. For our analysis purpose I have divided our target variable HUIDCOG into a binary response where 1 refers to the patient is healthy in terms of cognitive abilities and 0 is unhealthy. We assign 1 if the patient is able to think clearly and solve day to day problems (COG. ATT. LEVE 1) and assign a 0 otherwise since in any other case it indicates some kind of issues with the patient's cognitive health abilities.

```
hui$HUIDCOG = ifelse(hui$HUIDCOG=='COG. ATT. LEVE 1', 1, 0)
```

Dependent Variable Removal:

Next we are interested in seeing the different class distribution of the 8 dependent variables. Particularly the variables HUIGDEX and HUIGSPE.

```
## $HUIGDEX
## freq_dist
##    LIM. HANDS/F      NOT STATED USE OF HANDS/F.
##    0.0125644491    0.0001501727    0.9872853782
##
## $HUIGSPE
## freq_dist
##      NO PROBLEMS       NOT STATED PARTIAL/NOT UND.
##     0.9922410772     0.0002002303     0.0075586925
```

Both of these variables, in my opinion, lacks diversity and is biased heavily towards one class than the others, therefore I have decided not to include these 2 variables in my analysis.

```
hui = select(hui, -c(HUIGDEX, HUIGSPE))
```

Dependent Variable Group Collapsing:

Furthermore, I have tried to reduce the number of groups for each of the remaining 6 variables to some degree where it makes sense. For example, instead of using 4 possible classes of HUIGHER, which classifies the respondents based on their hearing state, I have collapsed it into 2, marking HUIGHER as Good if there was a history of hearing problem (whether or not its corrected currently) or otherwise Bad there was no hearing complains ever for that particular patient. Below is a summary of the collapsing decisions that has been made in this analysis.

HUIDEMO : Emotional index

This variable classifies respondents based on emotional health status. The original record has 6 different levels based on different levels of emotional response. But we can reduce it to Happy or Unhappy based on broader definition. I have converted all of NOT STATED as Unhappy.

Table 1: Mapping Table of HUIDEMO

HUIDEMO	isHappy
EMOT. ATT. LEV.1	Happy
EMOT. ATT. LEV.2	Happy
EMOT. ATT. LEV.3	Unhappy
EMOT. ATT. LEV.4	Unhappy
EMOT. ATT. LEV.5	Unhappy
NOT STATED	Unhappy

HUIGHER : Hearing State

This variable classifies respondents based on hearing state of the patient. As explained earlier, the original 4 possible classes are reduced to a broader 2 general classes of Good or Bad indicating if the patient has a history of hearing issues.

Table 2: Mapping Table of HUIGHER

HUIGHER	hearingState
NO PROBLEMS	Good
NOT STATED	Bad
PROB./CORR.	Good
PROB./NOT CORR.	Bad

HUIGMOB : Mobility Trouble

This variable classifies the respondents based on their state of mobility trouble. We classify this as TRUE or FALSE indicating if the respondent indicated that (s)he cannot move freely without external help.

Table 3: Mapping Table of HUIGMOB

HUIGMOB	mobilityHelp
NEED MECH. SUPP	TRUE
NO AID REQUIRED	FALSE
NO PROBLEMS	FALSE
NOT STATED	FALSE
REQUIRES HELP	TRUE

HUIGVIS : Vision State

This variable classifies the respondents based on their vision state. Like HUIGMOB, I have mapped this to TRUE if the respondent has a history of vision problem and False otherwise.

Table 4: Mapping Table of HUIGVIS

HUIGVIS	visualProb
NO PROBLEMS	TRUE
NOT STATED	TRUE
VISUAL P. UNCOR.	FALSE
VISUAL PROB. COR	FALSE

DHHGAGE : Age

Instead of age groups, I have take the mean age of the group, although since I am not reducing the number of classes here, it is probably not going to add a lot of value in the complexity reduction of our final model, unless we decide not to use this variable.

Table 5: Mapping Table of DHHGAGE

DHHGAGE	meanAges
45 TO 49 YEARS	47
50 TO 54 YEARS	52
55 TO 59 YEARS	57
60 TO 64 YEARS	62
65 TO 69 YEARS	67
70 TO 74 YEARS	72
75 TO 79 YEARS	77
80 TO 84 YEARS	82
85 AND OLDER	87

Here is the glance of the final dataset after data cleaning that we will be using in our model building.

Table 6: Final Dataset To Be Used For Model Building

DHH_SEX	HUIDCOG	isHappy	hearingState	mobilityHelp	visualProb	meanAges
MALE	1	Happy	Good	FALSE	TRUE	47
MALE	1	Happy	Good	FALSE	TRUE	47
FEMALE	1	Happy	Good	FALSE	FALSE	47
MALE	1	Happy	Good	FALSE	TRUE	47

DHH_SEX	HUIDCOG	isHappy	hearingState	mobilityHelp	visualProb	meanAges
MALE	1	Happy	Good	FALSE	FALSE	47
FEMALE	1	Happy	Good	FALSE	TRUE	47

```
##    DHH_SEX         HUIDCOG          isHappy      hearingState
##  FEMALE:11372   Min.   :0.0000   Happy  :18970   Bad : 1073
##  MALE  : 8605   1st Qu.:0.0000   Unhappy: 1007   Good:18904
##                 Median :1.0000
##                 Mean   :0.6983
##                 3rd Qu.:1.0000
##                 Max.   :1.0000
##  mobilityHelp    visualProb        meanAges
##  Mode :logical   Mode :logical   Min.   :47.00
##  FALSE:17816     FALSE:15638     1st Qu.:57.00
##  TRUE :2161      TRUE :4339      Median :67.00
##  NA's :0         NA's :0         Mean   :66.96
##                                  3rd Qu.:77.00
##                                  Max.   :87.00
```

2 Methods

In this part of the project, I have used some of the classification techniques that were taught over the course from Logistic Regression to Support Vector Machines. However for focussing on one technique, **Logistic Regression** is preferred which is interpretable and gives a low missclassification error rate as well as decent Specificity and Sensitivity score. I have used the variable importance table from random forest models to select a subgroup of variables to further tune my models. I have attached the details into the **Appendix** section.

2.1 Logistic Regression

Logistic Regression uses the logistic function fitted by **maximum likelihood**. It performs well even if the predictors do not follow Gaussian distribution. The model is a linear model in the log-odds of success

$$\log\left(\frac{p(X)}{1-p(X)}\right) = X\beta = \beta_0 + X_1\beta_1 + \ldots + X_p\beta_p.$$

Since our dependent variable takes a 0/1 binary response, we can use this model. Unlike linear regression where one unit change in the predictor variable (X) results in one unit change in Y, here one unit increase in X_j , while holding all others fixed is associated with a β_j change in the log-odds.

Let's start with baseline model in logistic regression. Before fitting the model, dataset is split into training and testing set in random sampled fashion of ratio 70:30. The model is fitted to train data and then predict the test data to validate based on its accuracy, sensitivity, specificity, etc.

The coefficients must be estimated based on the available training data. For logistic regression, the more general method of maximum likelihood is preferred for its robust statistical properties. Basically, the algorithm tries to find coefficients that maximize the likelihood that the probabilities are closest to 1 for people who don't have any problem in terms of patient's cognitive abilities (i.e. the respondent is able to think clearly and can solve day to day problems), and close to zero for people who has some type of cognitive disability and cannot carry out their day-to-day activity without some degree of help. During my experiment I have found that isHappy, mobilityHelp, hearingState, and meanAges are the most important variables so I have included only these 4 variables in my final model. Below table shows summary of estimates, Std.Error and p-value in the order of significance after performing logistic regression to the training data.

Table 7: Summary of Final Logistic Regression and its odd-ratios

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.58051491889384	0.11923118860264	13.2558849527299	4.17e-40
isHappyUnhappy	-1.25415543187966	0.0682410074257045	-18.3783252796362	1.96e-75
mobilityHelpTRUE	-0.524639323044865	0.0506502883766866	-10.3580717871361	3.85e-25
hearingStateGood	0.577738692439939	0.0661056790755542	8.73962268475639	2.34e-18
meanAges	-0.0169982650062456	0.00137517800999301	-12.3607743017445	4.26e-35

And here is the confusion matrix for the model indicating various measurements including accuracy, its 95% confidence interval, sensitivity, specificity and balanced accuracy as well.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0  1607  4421
##          1  1608 12341
##
##                Accuracy : 0.6982
##                  95% CI : (0.6918, 0.7046)
##     No Information Rate : 0.8391
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1744
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.49984
##             Specificity : 0.73625
##          Pos Pred Value : 0.26659
##          Neg Pred Value : 0.88472
##              Prevalence : 0.16094
##          Detection Rate : 0.08044
##    Detection Prevalence : 0.30175
##       Balanced Accuracy : 0.61805
##
##        'Positive' Class : 0
##
```

3. Results

3.1 Model Interpretation

On examining the fitted logistic regression model summary above, we can see that all the predictors are **statistically significant** with p-values far less than required 5%. These results also concur with the findings that predictors such as HUIGDEX, HUIGSPE, and DHH_SEX are not very significant predictors in determining a respondents cognitive health state. When we compare the full model of all available variables *(DHH_SEX + meanAges + isHappy + hearingState + mobilityHelp + visualProb)* with that of a model build on the subset of 4 predictors ($isHappy + mobilityHelp + hearingState + meanAges$) using Anova chisq test, a small p-value $(7.693e-08) < 0.05$ indicated that both models are similar - thus parsimoniously we chose the smaller model. Also, we have checked with various available logistic regression model selection techniques such as best subset, forward, backward and step-wise selection, all of them pointed to the four

predictors that was used in the final model build. Therefore the final model that we've selected for predicting **HUIDCOG** is:

$$y_i = \beta_0 + isHappy * \beta_1 + mobilityHelp * \beta_2 + hearingState * \beta_3 + meanAges * \beta_4$$

We can interprete the model in this way: If a respondent reported that (s)he requires mobility help == 1, keeping other predictor unchanged, that can be associated with an estimated increase of (-0.53) units in the log-odds of the respondent being cognitively healthy. We can see from the Table 7 that it is estimated that if the respondent is classified as happy or need no mobility help or has a good hearing state or is younger (selecting any one of these predictor while not changing the others) it generally is an indication of the respondent is cognitively healthy, which makes sense. Now we can not only point out which predictor is associated with diminishing cognitive ability but we can also indicate more relevant statistics which is by how much units they affect the mental state of the patients.

3.2 Model Evaluation

Logistic regression model is used to predict the test data as well as the validation dataset that was released later to the students. We have used various statistical measures to measure the effectiveness of this model such as missclassification rate, sensitivity, specificity and Area under the ROC curve. We have also tried k fold cross validation on the training data set. Below is a graph that shows the ROC curve plot that is derived on the test dataset. The value of Area under the ROC curve we got is ~0.61. Based on various cut-off values, we found a cut-off of 0.5 leads to the best balance of accuracy, sensitivity and specificity. Please refer to the Appendix for further details on this section's derivations.

3.3 Comparisons of Classification Models

This sections shows the various models that we tried along with Logistic Regression model that was ultimately selected. The classification was model was build on 70% of the available data and 30% of the data was used for testing. Different metric that are used to compare the models are discussed in the following:

Mis-classification Error: The number of observations that were predicted wrongly by the model. It is the proportion of misclassified observations.

Sensitivity: It is the ability of a model to correctly identify those with diabetic disease. It is observed True positive rate. TP/(TP+FN) where TP is True Positive and FN is False Negative.

Specificity: It is the ability of a model to correctly identify those without diabetic disease. It is observed True negative rate. TN/(TN+FP) where TP is True Negative and FP is False Positive.

Area Under Curve (AUC): It is a measure of the overall performance of the classifier. If it is close to one, then model is performing as good as the real truth. On the other hand, if AUC is 0.5, then model would perform no better than null classifier.

```
# ToDO different models
```

Conclusion and Discussion

After going through all the models, we have picked the logistic regression model with four predictors as our final selected model. However we have only considered linear terms in this model, which can be a drawback in this model. Therefore in the future it will be noteworthy to check how adding interaction and non-linear terms changes the predictive ability of the logistic regression model, or if there are any other family of classification techniques stands out when these new terms are introduced. We can try for even more powerful models such as deep neural networks to see if that helps us to give a better result. Finally all these discussions are based on the dataset that was provided to us. The correctness of this model will change as new data are available to us in the future, so that we might have to tune it later.