Classification of the AUDITC Among the General US Population

Adam J Batten

Spring 2014

Abstract

Alcohol studies in the US Veteran population to date have shown some interesting results regarding the self-identified quantity of current alcohol consumption. Namely, those classified as having never consumed alcohol are shown to have increased risk of several comorbid conditions including chronic vascular disease. To date few studies have delved into the accuracy of the alcohol questionnaire nor have their been any attempts to develop a more accurate prediction model of alcohol consumption. Using the National Health Interview Survey Adult Sample (NHIS) data for year 2012 we develop 2 prediction models for classifying subjects within the categories of the AUDITC[9]. The first model, a cross-validated multinomial lasso model resulted in a 45% flat classification error rate (10853 cross-entropy error) when predicting categories of the AUDITC. We also tuned a 2 layer neural net with standard back-propagation which resulted in a 28% flat error rate (10477 cross-entropy error). While the neural net is clearly a more accurate classifier, more work needs to be done assessing the failure of both models to accurately separate the non drinking and heavy drinking classes of the AUDITC.

1 Introduction

The National Health Interview Survey (NHIS)[8] is one of the largest public health surveys on the planet. Run yearly since 1957 it is a rich source for health information on the adult (age 18 and over) US public. The survey instrument covers the most basic information (gender, age, race, etc.) as well as some very detailed health information (diagnosis of comorbid conditions, mental health status, alcohol consumption, etc.). In analyzing this data researchers must be careful to adjust for the multi-stage stratified survey methodology of the NHIS which over-samples asian, black, and hispanic populations.

In defining our feature set we wanted to exclude questions that were new and experimental. Of the 881 possible features in the 2012 NHIS, we identified 590 that were included in years 2009-2011. However, many of these are redundant or include questions that the NHIS admits may be confusing for respondents. For example there are 17 questions measuring which joint causes the respondent discomfort. We have collapsed features such as these into single features (see attached R code for details). After excluding features missing from at least 1 year of the survey between 2009 and 2012 and either collapsing or combining redundant features our final prediction space includes 79 features. We include information on comorbidities such as CVD, diabetes, depression, and COPD; socio-economic status information such as employment status (unemployed, retired, student, etc...), occupation type, and some classic instrument variables such as whether or not the respondents' current job includes sick leave benefits. We also include race and ethnicity information, which is gathered according to the definitions set forth by the Office of Management and Budget. In addition to these demographic, comorbid, and SES information we also collect information on the respondents' perceived health (both mental and physical), their activity level, and some additional information on functional limitations. It should be mentioned that some variables are top coded to protect respondent identity; i.e. weight is bottom and top coded at 100 and 299 pounds respectively. Similarly height is restricted to 59-76 inches, BMI between 14 and 56, and all age features are top-coded at 85.

We use the standard AUDITC definition from the Veterans Health Administration[9] which combines a set of 11 questions pertaining to the respondents current alcohol consumption into a single factor variable with 12 levels. Generally a male who scores above 4 and a female who scores above 3 on the AUDITC is considered to be at risk for worse health outcomes. It is standard practice to aggregate the AUDITC into a 3 category factor (None, Moderate, and Heavy) where None represents those respondents who are currently not drinking or are lifetime abstainers. Missing responses are often dropped from analysis but for this paper we assign them a special missing category.

2 Training Set

The training set comprised 2/3 (N=18,480) of the 2012 NHIS survey respondents. Most of the features are distributed similarly across the levels of AUDITC with the bulk of the population centered within the Moderate to Heavy drinking categories (see table 1). Whites, men, married or people with domestic partnerships, and employed respondents are less likely to identify themselves as non-drinkers. Comorbidities are also distributed similarly across classes of AUDITC, with slightly higher rates in the Moderate Drinking category.

Table 1: Demographics within levels of AUDITC.

	None	Moderate	Heavy	Missing
Male	0.07	0.18	0.15	0.08
Female	0.07	0.19	0.12	0.14
White	0.12	0.29	0.24	0.16
Black	0.02	0.04	0.02	0.04
AIAN	0.00	0.00	0.00	0.00
Asian	0.00	0.02	0.01	0.02
Secret	0.00	0.00	0.00	0.00
Multi	0.00	0.01	0.01	0.00
NotHispanic	0.12	0.31	0.24	0.17
Hispanic	0.02	0.05	0.03	0.05
Divorced	0.02	0.04	0.03	0.01
Partner	0.09	0.23	0.16	0.11
Separated	0.01	0.01	0.01	0.01
Single	0.02	0.07	0.07	0.07
Unknown	0.01	0.02	0.01	0.02
Unemployed	0.08	0.12	0.08	0.11
Employed	0.06	0.25	0.19	0.11
Northeast	0.02	0.07	0.05	0.04
Midwest	0.03	0.09	0.07	0.04
South	0.06	0.12	0.09	0.09
West	0.03	0.08	0.06	0.05

Table 2: Comorbid conditions within levels of AUDITC.

		None	Moderate	Heavy	Missing
Hypertension	No	0.08	0.27	0.20	0.16
	Yes	0.06	0.10	0.07	0.07
CHD	No	0.13	0.35	0.26	0.21
	Yes	0.01	0.01	0.01	0.01
Angina	No	0.14	0.36	0.27	0.22
	Yes	0.01	0.01	0.00	0.01
MI	No	0.13	0.35	0.27	0.21
	Yes	0.01	0.01	0.01	0.01
Heart Disease	No	0.12	0.34	0.26	0.21
	Yes	0.02	0.02	0.01	0.01
Stroke	No	0.13	0.36	0.27	0.21
	Yes	0.01	0.01	0.00	0.01
Emphysema	No	0.13	0.36	0.27	0.22
	Yes	0.01	0.01	0.00	0.00
Asthma	No	0.12	0.32	0.24	0.20
	Yes	0.02	0.05	0.03	0.03
Ulcer	No	0.13	0.34	0.26	0.21
	Yes	0.02	0.02	0.01	0.01
Cancer	No	0.12	0.33	0.25	0.21
	Yes	0.02	0.03	0.02	0.02
Diabetes	No	0.12	0.33	0.26	0.20
	Yes	0.02	0.03	0.01	0.03

3 Methodology

We build two models to predict the AUDITC classification among the general US public, a model based on the multinomial Lasso regression of Tibshirani[2] and a 2 layer Neural Net standard back-propagation model of Rumelhart[3]. We use the R language for statistical computing with packages glmnet[6] for the lasso and RSNNS[7] for the neural net. When tuning our lasso model we carefully conduct group sensitive cross-validation by ensuring no respondents from the same household are separated during the training process.

For our lasso model, using 10-fold cross validation and the default lambda sequence[6] we obtained a minimal cross-entropy error of 21716 (0.45 flat error loss) for lambda=0.0007. The confusion matrix for this tuning set is displayed below and reveals the relatively poor ability of the model to accurately classify both the central classes (Moderate and Heavy) and the boundary classes (None and Missing). This is especially true of the Moderate drinking class which has a high number of respondents misclassified as Missing. However, it is encouraging to see the ability of the model to classify the Missing category with such precision.

Table 3: Lasso CV confusion matrix.

	Predicted			
AUDITC	None	Moderate	Heavy	Missing
None	640	1250	378	654
Moderate	351	3832	1220	1062
Heavy	195	2329	1764	511
Missing	267	1580	311	2136

For our 2 layer neural net model we tune our classification model using 3 values for each layer and learning rate. Computational limitations and time constraints prevented us from conducting cross validation along with the tuning on our neural net. The tuning process reveled a minimal cross-entropy error of 23307 (0.45 flat error loss) with a learning rate of 0.01, layer 1 node size of 10, and layer 2 node size of 8.

Table 4: Neural Net CV tuning grid and cross-entropy prediction error.

	Size1	Size2	Learning Rate	CE Error
1	10	3	0.01	26091.97
2	15	3	0.01	24444.99
3	20	3	0.01	25330.45
4	10	5	0.01	25683.11
5	15	5	0.01	26315.79
6	20	5	0.01	25640.23
7	10	8	0.01	23306.94
8	15	8	0.01	24594.13
9	20	8	0.01	26849.61
10	10	3	0.10	26169.36
11	15	3	0.10	25008.19
12	20	3	0.10	25609.69
13	10	5	0.10	25524.21
14	15	5	0.10	24801.89
15	20	5	0.10	26837.64
16	10	8	0.10	25210.53
17	15	8	0.10	25632.53
18	20	8	0.10	24447.59

The neural net confusion matrix for this tuning set is displayed below and though it is more accurate than the lasso results, the neural net also has some difficulty separating the Moderate and Heavy categories.

Interestingly the neural net is only slightly better able to classify the boundary classes (None and Missing).

Table 5: Neural Net CV confusion matrix

	Predicted			
AUDITC	None	Moderate	Heavy	Missing
None	878	405	250	310
Moderate	1159	3967	1928	1333
Heavy	391	1314	2273	334
Missing	494	779	348	2317

4 Test Set and Results

The test set comprised 1/3 (N=9251) of the 2012 NHIS survey respondents. Distributions across AUDITC categories are similar to those of the training set.

Table 6: Test Set demographics within levels of AUDITC.

	None	Moderate	Heavy	Missing
Male	0.07	0.18	0.15	0.08
Female	0.08	0.17	0.12	0.15
White	0.12	0.29	0.23	0.16
Black	0.02	0.04	0.02	0.04
AIAN	0.00	0.00	0.00	0.00
Asian	0.01	0.02	0.01	0.02
Secret	0.00	0.00	0.00	0.00
Multi	0.00	0.01	0.00	0.00
NotHispanic	0.12	0.30	0.23	0.18
Hispanic	0.02	0.05	0.03	0.05
Divorced	0.02	0.03	0.02	0.01
Partner	0.09	0.23	0.16	0.12
Separated	0.00	0.01	0.01	0.01
Single	0.02	0.07	0.07	0.07
Unknown	0.01	0.02	0.01	0.02
Unemployed	0.08	0.12	0.08	0.12
Employed	0.07	0.24	0.19	0.11
Northeast	0.02	0.07	0.05	0.04
Midwest	0.03	0.09	0.07	0.04
South	0.06	0.12	0.09	0.09
West	0.03	0.08	0.07	0.05

A heat map of the feature relationships within the test data reveals that there are few correlated features.

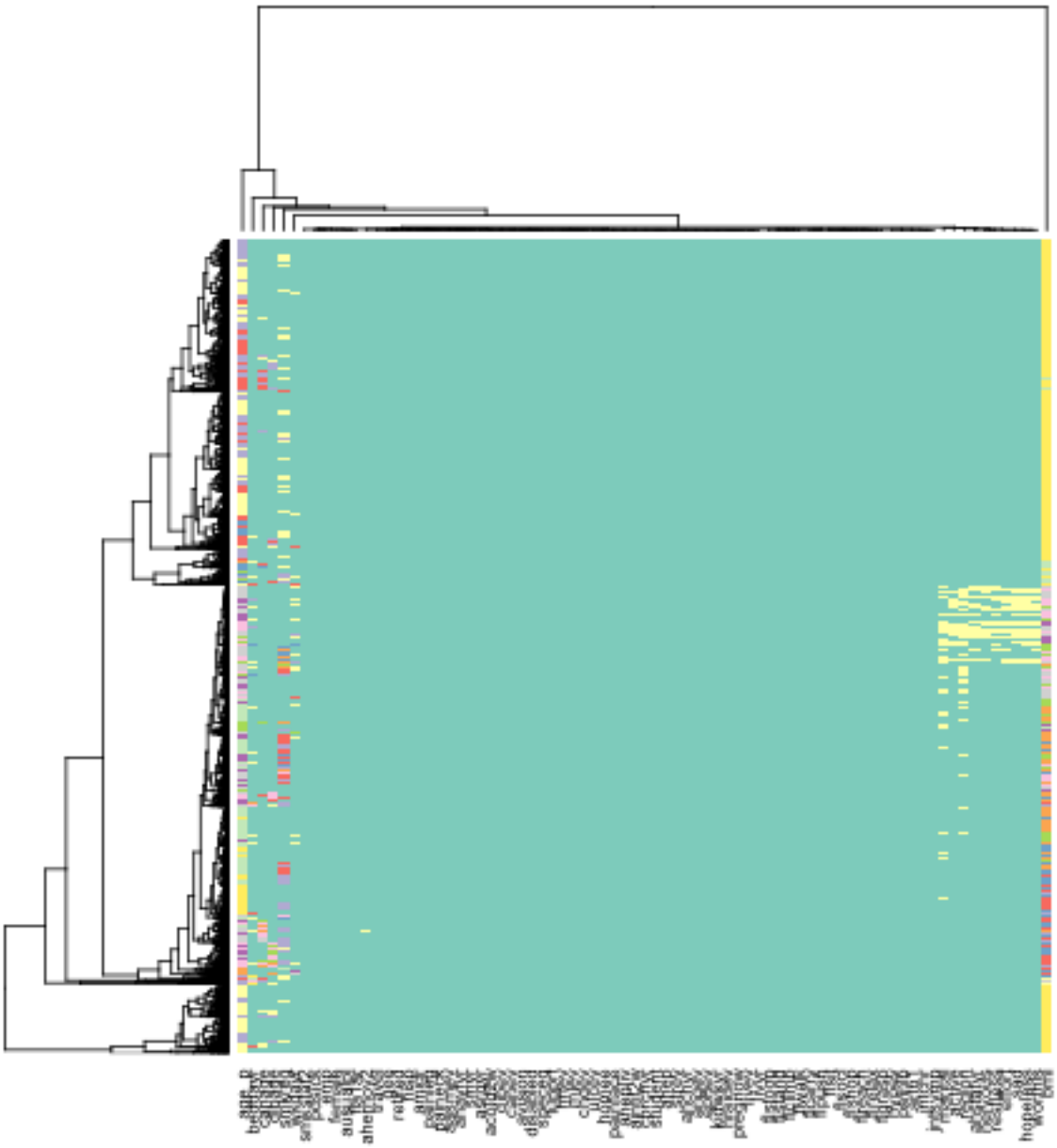


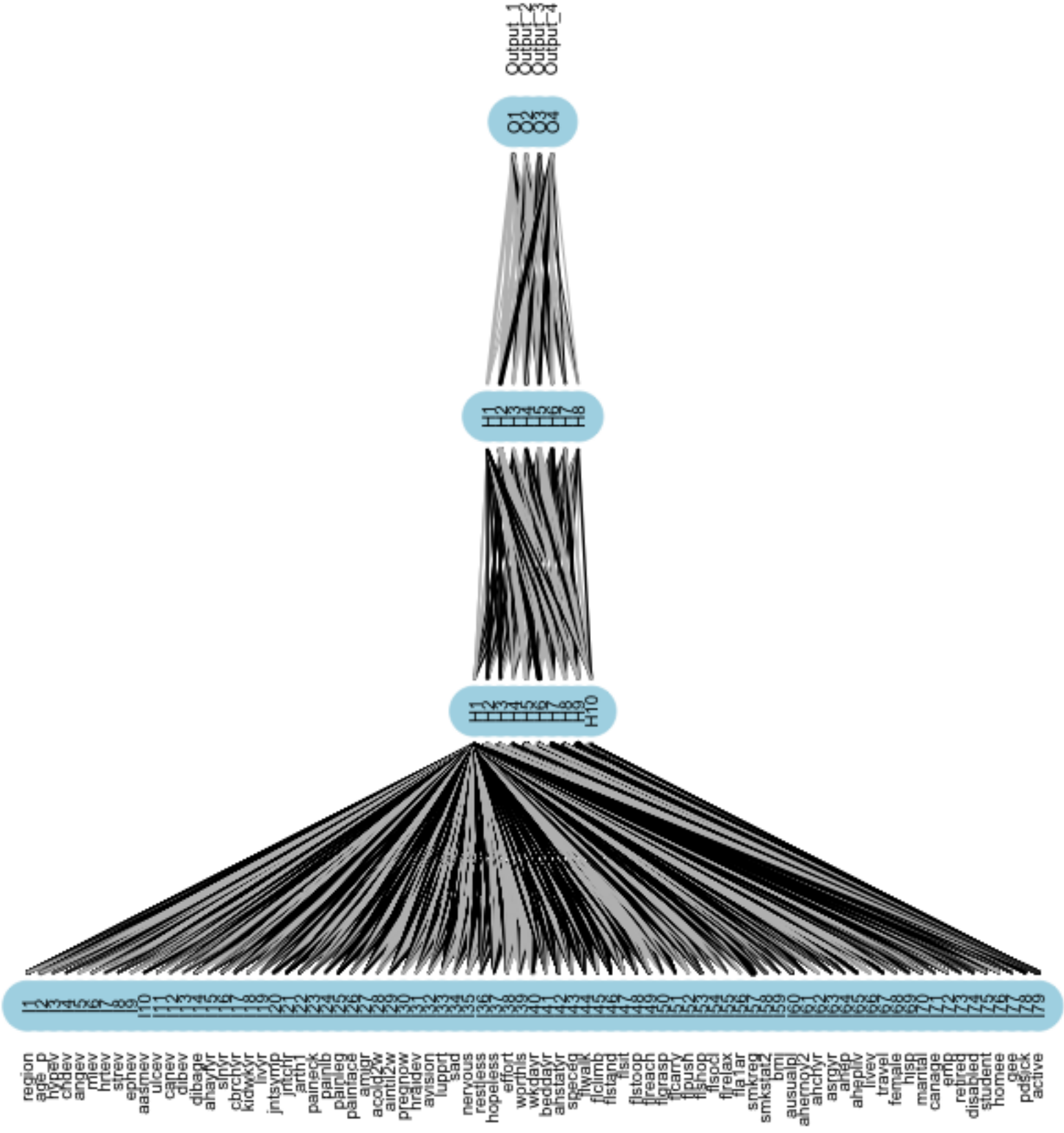
Figure 1: Heatmap of the NHIS Training Set features

Using the tuned models, our final prediction cross entropy error for the lasso was 10853 (0.45 flat error) and 10477 (0.28 flat error) for the 2 layer neural net. The neural net leads to a reduction in cross entropy error of 3.5% over the lasso. Looking closer at the final confusion matrix from the neural net prediction we see it has reduced the number of Moderate drinkers misclassified as Missing, but there remains some difficulty in separating the Moderate and Heavy drinkers.

Table 7: Two layer neural net confusion matrix for AUDITC classification

	Predicted			
AUDITC	None	Moderate	Heavy	Missing
	None	Moderate	Heavy	Missing
None	317	265	135	222
Moderate	590	1632	1047	740
Heavy	212	772	1004	215
Missing	316	525	237	1022

A plot of the hidden layers in the neural network reveals the relationships between features. Output_1 corresponds to AUDITC level None, Output_2 to Moderate, and so on. The darker lines correspond to heavier associations, for example the node H2 in the second layer is highly associated with the Missing AUDITC category.


region
age_p
hypev
chdev
angev
miev
hrtev
strev
ephev
aasmev
ulcev
canev
dibev
dibage
ahaytyr
sinyr
cbrchyr
kidwkyr
livyr
intswmp
jntchr
arth1
paineck
painlb
painleg
painface
amgr
acold2w
aintl2w
pregnow
hraidev
anvision
luppn
sad
nervous
restless
hopeless
effort
worthls
wkdayr
beddayr
ahstatyr
speced
flwalk
flclimb
flstand
flsit
flstoop
flreach
flgrasp
flcarry
flpush
flshop
flsocl
flrelax
fla1ar
smkreg
smkstat2
bmi
ausualpl
ahernoy2
ahchyr
asgyr
ahep
ahepliv
livev
travel
female
hisp
marital
canage
emp
retired
disabled
student
homee
gee
pdsick
active
H10
H8
output_1
output_2
output_3
output_4

If we look closer at the node weights for the first layer we see some interesting feature relationships. For example the first node of layer 1 (H1) is more associated with features including angina (angev), liver problems (liver), difficulty sitting, being employed, a student, and a home engineer. Additionally some of the nodes are highly specialized, like node 5 (H5) which is most associated with major comorbidities (chronic heart disease, myocardial infarction, cancer) and ER visits (ahernoy2). There are also some curious nodes like node 7 which includes currently pregnant, liver disease, paid sick leave, and activity level. The CVD features are split between nodes H1 with angina; H5 with chronic heart disease and myocardial infarction; H6 with hypertension and stroke; and H9 with any non specific heart condition (hrtev).

Table 8: Layer 1 feature assignment

L1 Node	Features
H1	angev, livyr, flsit, ahchyr, emp, student, homee
H2	age, cbrchyr, hraidev, nervous, effort, flshop, flsocl
H3	region, ephev, ulcev, dibage, sinyr, painleg, lupprt, restless, beddayr, flreach, flcarry, flpush, gee
H4	ahstatyr, fla1ar
H5	chdev, miev, canev, ahernoy2
H6	hypev, strev, painlb, painface, wkdayr, speceq, flclimb, flgrasp, smkstat2, ahepliv, female, canage, disabled
H7	paineck, acold2w, pregnow, avision, flstand, ausualpl, livev, pdsick, active
H8	aasmev, kidwkyr, jntsymp, arth1, amigr, sad, hopeless, flstoop, asrgyr, marital
H9	hrtev, dibev, jntchr, aintil2w, worthls, flwalk, flrelax, smkreg, bmi, ahep, travel, hisp, retired
H10	ahayfyr

Looking at the relationship between the 2 layers and the AUDITC classification we see that all but 2 of the L1 nodes associated with CVD features (H6's hypertension and stroke; H9's any heart condition) are more associated with L2 node 6 (HH6) which in turn is strongly associated with both the non drinking and the heavy drinking categories of the AUDITC.

Table 9: Relationship between hidden layers

	L2 Node							
L1 Node	HH1	HH2	HH3	HH4	HH5	HH6	HH7	HH8
H1	-0.53	-1.31	0.24	0.90	-0.33	1.61	-1.17	0.60
H2	0.44	1.09	0.42	-2.42	1.97	-0.85	0.40	-0.98
H3	0.50	1.05	0.89	-1.80	1.78	-1.29	0.47	-1.35
H4	0.25	-0.78	0.18	0.53	-2.19	-0.75	-0.06	0.10
H5	0.07	-2.56	-0.31	-0.09	-0.45	2.40	-0.40	0.81
H6	-0.30	2.90	0.91	-1.05	1.86	-1.46	-0.07	-1.26
H7	-0.33	-2.49	1.30	0.68	-0.85	1.74	0.45	0.42
H8	0.38	-2.07	-0.46	0.92	-2.01	0.42	-0.31	0.52
H9	1.07	-2.04	0.08	-0.28	-1.30	0.85	1.92	-0.08
H10	0.30	-0.83	-1.69	1.59	-2.59	0.18	-1.31	0.78

	AUDITC			
Layer 2 Node	None	Moderate	Heavy	Missing
HH1	-1.81	-0.41	-0.30	-0.23
HH2	-2.48	-2.19	-1.09	2.33
HH3	-0.06	0.90	-1.57	-0.94
HH4	-3.17	-0.72	2.08	-0.59
HH5	1.64	-2.74	-0.85	1.56
HH6	1.43	-0.66	1.33	-1.86
HH7	-1.00	1.42	-0.91	0.29
HH8	-0.37	-1.38	0.11	-0.38

Interestingly though L1 node 6 is strongly associated with the same L2 node 6, it is more strongly associated with L2 node 2 which is most associated with the missing category of the AUDITC. Looking at some of the other features in L1 node 6 we see chronic joint pain (jntchr), intestinal illness in the last 2 weeks (aintil2w), feeling worthless (worthless), difficulty walking (flwalk), difficulty relaxing (flrelax), former smoker (smkreg), BMI, hepatitis (ahep), foreign travel (travel), hispanic, and retired. These features tell an interesting story as many of them are confounded with alcohol consumption. There is some concern that the missing categorization might actually not be missing at random as it could very well be the case that these respondents are too sick (either temporarily with a stomach bug or permanently with gout) to be drinking. That many of the same features associated with the non drinking AUDITC category are also associated with the heavy drinking category is reflective of research to date in the VA population on the accuracy of self reported AUDITC scores[1].

5 Discussion

Both methods have difficulty separating the Moderate and Heavy drinkers and this could be an artifact of the AUDITC definition which may be too coarse a measure to separate those that are truly moderate drinkers from moderate drinkers who occasionally binge drink. Though the neural network has challenges classifying the outer classes, this may be an artifact of the inability of the multilayer network to account for the NHIS survey weights. Though [5] has tackled the survey weighting problem for the simple one layer case, there is no general solution for the multilayer perceptron. One potential solution to this problem that is beyond the scope of this paper comes from [4] who recommends using local mixtures of experts within the strata of complex surveys and weighting each expert's predictions by the survey weights. To increase the accuracy of the neural net we thought to include additional hidden layers or even expand the sample to include prior year surveys.

References

[1] Delaney

[2] LASSO

[3] Rumelhart

[4] Amer

[5] Montanari 2005

[6] Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL glmnet

[7] Christoph Bergmeir, Jose M. Benitez (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. Journal of Statistical Software, 46(7), 1-26. URL RSNNS

[8] National Health Interview Survey Center for Disease Control and Prevention Survey Description 2012 NHIS 2012

[9] AUDITC