DSC 478
Michal Chowaniak
Final Project Report

Executive Summary

The 2017 National Health Survey adult data set was used alongside decision tree and logistic regression analytical models in attempt to answer a question about what kind of patients are told to exercise more. To author surprise none of disease, demographic, employment related variables were chosen by the model. However, the lesson learned was very straightforward, patients were told about need of exercising more when they were not told to reduce fat/calories in diet, were not increasing physical activity, were not told to participate in weight loss program, were not seen health professional in last 6 months, or were not having any functional limitations. This could be a useful information for medical students and practitioners which want to specialize in obesity medicine.

Abstract

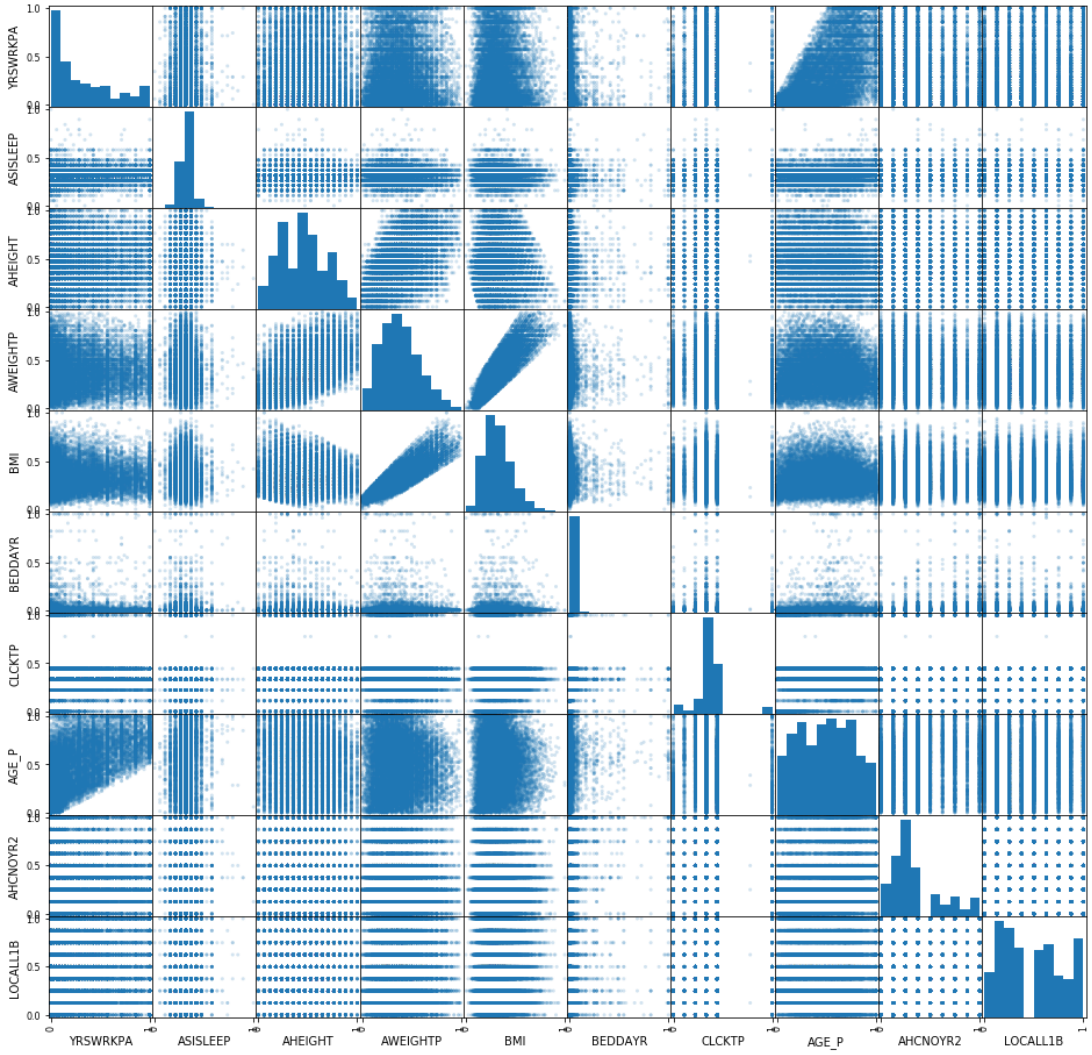
This project attempted to identify contributing factors for predicting what kind of patients are told to increase physical activity in past 12 months, using National Health Survey data made publicly available by The National Center for Health Statistics (NCHS). The analyses focused on investigating the significant predictors and interesting dependencies. The survey data set include mostly categorical variables, so the most appropriate methods were chosen, which were decision tree and logistic regression.

Introduction

The National Health Survey provides annual information on health status of the United States population, it was authorized by the National Health Survey Act in 1956. This project was focusing on “Adult” data set from the 2017 survey. Variables in the data set could be grouped to demographic, employment status, various illnesses. The data set included over 700 variables and over 30k observations. The data set had to be preprocessed, recoded and cleaned because it included a lot if missing values.

Methods and Results

Extensive preprocessing had to be performed in order to create a usable data set. Variables, which included thousands of missing values were dropped first. Next step was to bin several variables to make sure all variables were in the same range. Also, some values like Refused, Not ascertained, Don't know were recoded to missing values, so they could be removed later. All categorical variables were recoded to binary k-1 dummy variables. For some models numerical variables were normalized. Highly correlated numerical variables over 0.7 were dropped from the data set.

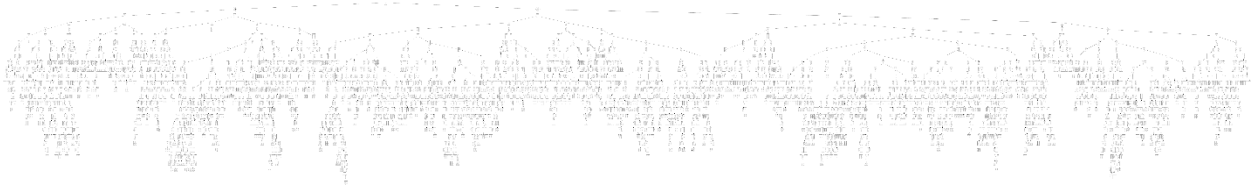


The cleaned data set included 142 variables and 15912 observations.

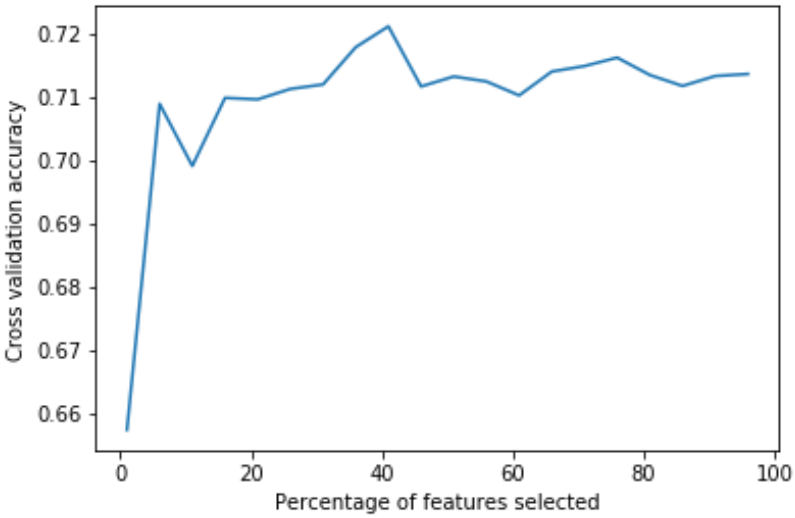
At first attempt to reduce dimension of the dataset a decision tree was run on train dataset which included 80% of observations

That model was 0.71 accurate, however tree produced was humongous. Further model tweaking was needed.

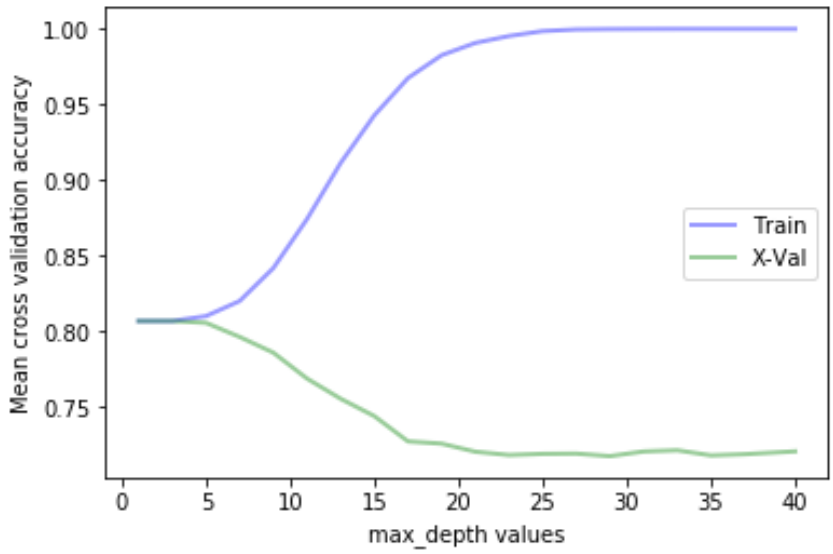
Next step was to run feature selection based on Chi squared test. Tree produced on 30% of the most significant variables was also very wide.



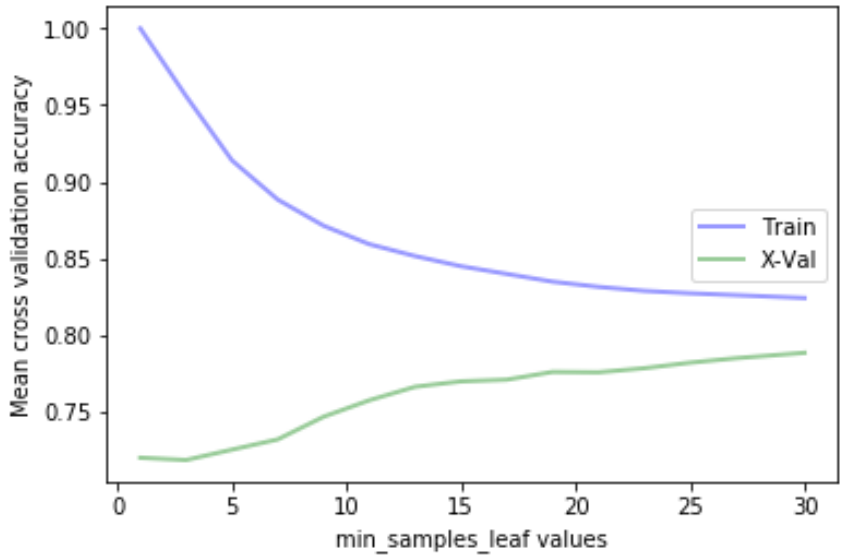
The best percentile of features using cross validation was 41, the accuracy was about 72% using 58 variables. However, 6 most significant features would give 0.71 accuracy.



The max depth using calc_params function returned following figure. Maximum depth was in range 4 to 5.



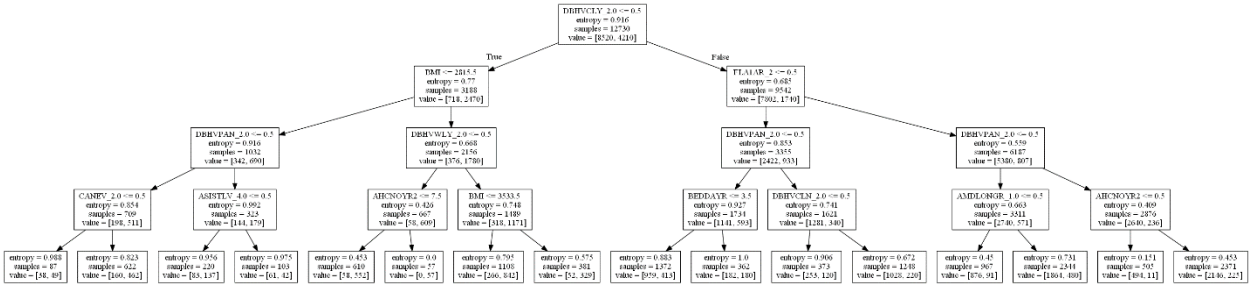
Min number of samples allowed at a leaf node equal to 10.



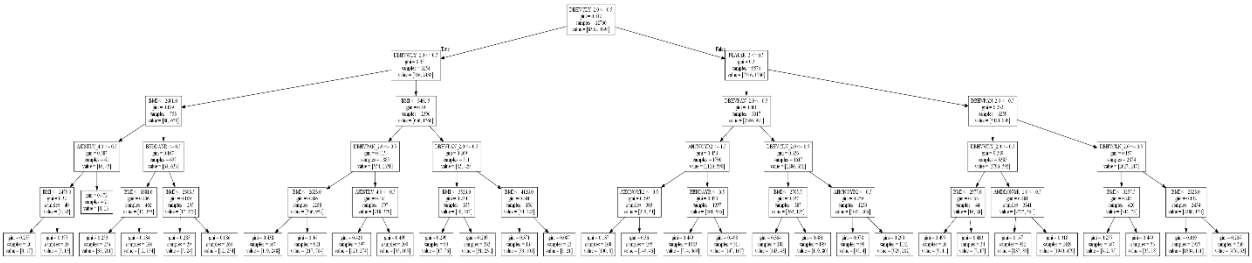
The decision tree run using above parameters returned accuracy of 0.803

Which was significant improvement.

```
Accuracy:0.803

Classification report
             precision    recall  f1-score   support

          0       0.81      0.92      0.86      2140
          1       0.78      0.56      0.65      1043

avg / total       0.80      0.80      0.79      3183


Confussion matrix
[[1973  167]
 [ 459  584]]
```



Another attempt to reduce dimensionality was to run decision tree using GridSearchCV with a combination of different parameters. The tree produced by this function increased accuracy to 0.806. However, it produced larger tree. It is worth to mention that the results were different almost each time this model was run.



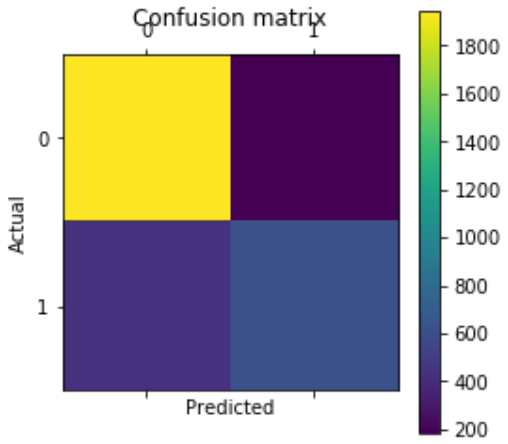
At the end variables chosen for logistic regression were the following”

DBHVCLY_2.0	Told to reduce fat/calories in diet, past 12 m: No
BMI	Body Mass Index
DBHVPAN_2.0	Currently increasing physical activity: No
DBHVWLY_2.0	Told to participate in weight loss program, past 12 m: No
CANEV_2.0	Ever told by a doctor you had cancer: No
ASISTLV_4.0	How worried are you about…maintaining standard of living: Not worried
AHCNOYR2	Total number of office visits, past 12 m
BEDDAYR	Number of bed days, past 12 months
DBHVCLN_2.0	Currently reducing fat/calories in diet: No
AMDLONGR_1.0	Time since last seen/talked to health professional: 6 months or less
AHCNOYR2	Total number of office visits, past 12 m
FLA1AR_2	Any functional limitation, all conditions: No

There were a few logistic regression models run, each of them produced slightly different results. Logistic Regression with Recursive Feature Elimination and Cross Validation returned all 14 variables as the most significant.

Logistic Regression Cross Validation with Recursive Feature Elimination returned 12 variables with 5 most significant feature as below:

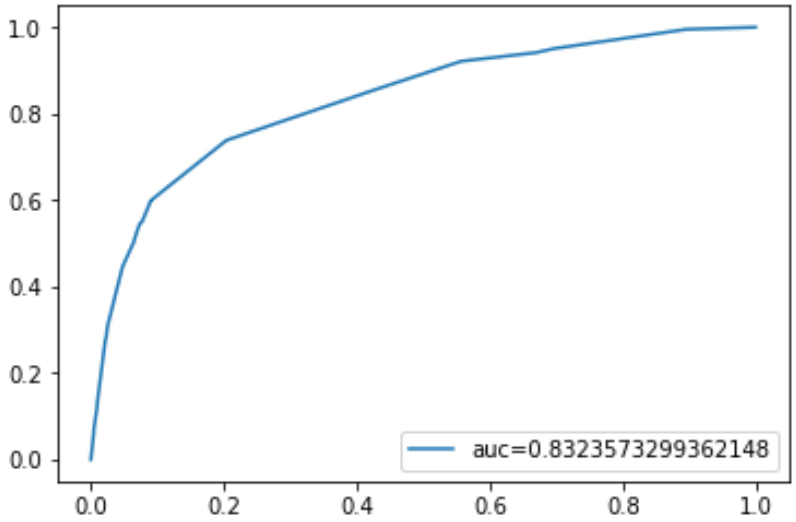
DBHVCLY_2.0	Told to reduce fat/calories in diet, past 12 m: No
DBHVPAN_2.0	Currently increasing physical activity: No
DBHVWLY_2.0	Told to participate in weight loss program, past 12 m: No
AMDLONGR_1.0	Time since last seen/talked to health professional: 6 months or less
FLA1AR_2	Any functional limitation, all conditions: No



Accuracy was 0.805, Precision 0.8 and recall 0.81

	precision	recall	f1-score	support
0	0.82	0.91	0.86	2128
1	0.77	0.59	0.67	1055
avg / total	0.80	0.81	0.80	3183

Area under curve was 0.83, which closer to 1 would be better.



To confirm above results and as a comparison, anther Logistic Regression was run using stasmodels.org API on full dataset. Variables with p-values less then 0.05 were manually removed.

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
SEX_2	0.2126	0.0658	3.2303	0.0012	0.0836	0.3415
R_MARITL_2	-0.1097	0.0834	-1.3152	0.1884	-0.2732	0.0538
R_MARITL_3	-0.1109	0.0627	-1.7698	0.0768	-0.2338	0.0119
R_MARITL_4	-0.0479	0.0604	-0.7935	0.4275	-0.1662	0.0704
MRACRPI2_2	0.0580	0.0780	0.7438	0.4570	-0.0949	0.2109
MRACRPI2_3	0.0799	0.2052	0.3892	0.6971	-0.3224	0.4821
MRACRPI2_4	0.6256	0.0970	6.4523	0.0000	0.4355	0.8156
REGION_2	0.0305	0.0703	0.4335	0.6646	-0.1073	0.1682
REGION_3	0.0828	0.0657	1.2607	0.2074	-0.0459	0.2116
REGION_4	0.1338	0.0705	1.8966	0.0579	-0.0045	0.2720
PAR_STAT_2	0.0067	0.1399	0.0482	0.9615	-0.2674	0.2809

The following most significant variables we chosen

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
SEX_2	0.1426	0.0476	2.9957	0.0027	0.0493	0.2360
MRACRPI2_4	0.5928	0.0921	6.4353	0.0000	0.4122	0.7733
WRKCATA_2.0	-0.2586	0.1106	-2.3386	0.0194	-0.4752	-0.0419
WRKCATA_5.0	-0.2313	0.0878	-2.6348	0.0084	-0.4033	-0.0592
HOURPDA_2.0	0.2143	0.0459	4.6726	0.0000	0.1244	0.3043
PDSICKA_2.0	-0.1553	0.0486	-3.1996	0.0014	-0.2505	-0.0602
HYPEV_2.0	-0.3228	0.0482	-6.7006	0.0000	-0.4172	-0.2284
HYBPLEV_5.0	0.4692	0.1426	3.2905	0.0010	0.1897	0.7487
CHLEV_2.0	-0.2340	0.0478	-4.8934	0.0000	-0.3277	-0.1403
CANEV_2.0	0.1573	0.0643	2.4450	0.0145	0.0312	0.2834
DBHVCLY_2.0	-2.2037	0.0524	-42.0296	0.0000	-2.3064	-2.1009
DBHVWLY_2.0	-1.0073	0.0988	-10.1991	0.0000	-1.2009	-0.8137
DBHVPAN_2.0	-0.7673	0.0487	-15.7697	0.0000	-0.8626	-0.6719
DBHVCLN_2.0	0.1892	0.0492	3.8479	0.0001	0.0928	0.2855
DBHVWLN_2.0	0.1701	0.0837	2.0309	0.0423	0.0059	0.3342
DIBREL_2.0	-0.1270	0.0450	-2.8227	0.0048	-0.2153	-0.0388
DIBPRE2_2.0	-0.3260	0.0706	-4.6195	0.0000	-0.4643	-0.1877
AHEARST1_2.0	0.1389	0.0449	3.0944	0.0020	0.0509	0.2268
AHEARST1_4.0	0.3072	0.1020	3.0107	0.0026	0.1072	0.5073
AHEARST1_6.0	0.8787	0.4135	2.1249	0.0336	0.0682	1.6891
VIMGLASS_2.0	-0.1661	0.0500	-3.3200	0.0009	-0.2642	-0.0681
AVISACT_2.0	0.3046	0.0523	5.8223	0.0000	0.2021	0.4072
FLA1AR_2	-0.4191	0.0480	-8.7321	0.0000	-0.5132	-0.3250
APLKIND_5.0	-0.7298	0.2222	-3.2837	0.0010	-1.1653	-0.2942
AMDLONGR_4.0	-1.0822	0.2259	-4.7913	0.0000	-1.5249	-0.6395
AMDLONGR_5.0	-1.3555	0.4590	-2.9530	0.0031	-2.2551	-0.4558
HIT1A_2.0	-0.1654	0.0525	-3.1518	0.0016	-0.2682	-0.0625
HIT4A_2.0	-0.2363	0.0573	-4.1268	0.0000	-0.3486	-0.1241
ASICPUSE_2.0	0.2521	0.0872	2.8907	0.0038	0.0812	0.4230
ASICPUSE_4.0	0.2588	0.0699	3.7014	0.0002	0.1218	0.3959
ASIMEDC_2.0	0.2205	0.0510	4.3265	0.0000	0.1206	0.3204
ASISTLV_4.0	-0.2352	0.0479	-4.9058	0.0000	-0.3291	-0.1412
AWEBUSE_2.0	0.1938	0.0802	2.4160	0.0157	0.0366	0.3510
YTQU_YG1_2.0	0.2135	0.0619	3.4489	0.0006	0.0922	0.3349
ASISLEEP	0.0322	0.0148	2.1674	0.0302	0.0031	0.0613
BMI	0.0006	0.0000	16.7943	0.0000	0.0005	0.0007
AHCNOYR2	0.0986	0.0104	9.4971	0.0000	0.0783	0.1190

Above model and variables significantly differ from the decision tree/logistic regression model developed manually. Similar accuracy was achieved with much larger number of variables.

The logistic regression analysis with decision tree as a preprocessing step for dimension reduction produced the model which can predict which patients are informed by healthcare professionals that they need to increase physical activity.

So which patients were told to increase physical activity in past 12 months? Those are only 5 the most significant results.

1. Those who were not told to reduce fat/calories in diet,
2. Those who were not increasing physical activity,
3. Those who did not participate in weight loss program,
4. Those who seen health professional in last 6 months,
5. Those who do not have any functional limitations,

Conclusion

The report analyzed adult data set for Health National Health Survey Data, the goal was to identify relationship between patients who were told to increase physical activity in past 12 months. Decision tree/LDA and logistic regression analysis were able to determine with 80% accuracy several predictors. However, the main limitation of this analysis is that different logistic regression models produced slightly different results. The next step in this research could be applying a different model in attempt to confirm results of logistic regression. The regression model results could be useful information for health care professionals, which specialize in obesity medicine. The predictors ended up being very different then were though at the beginning of the process. Notable, the very interesting observation was that none of disease related variables, none of demographic variables, none of employment related variables ended up in the model. This just shows that common sense gut feeling is not always correct.