Lasso-LDA-based Adult Autism Recognition (Preliminary Results Report)

Group name: Project 3

Group Member: Liuxi Wang (Student ID: 222025005)

Shule Ge (Student ID: 222025024)

Chenxing Liu (Student ID:222025026)

Chun Wang (Student ID:222025007)

Xiaoningsi Wang (Student ID:222025002)

Zhuang Chen (Student ID:222025021)

1. Process Overview

Topic Selection(100%) & Requirement Design(100%) => EDA(100%) => Variable Selection(100%)=> Data Clean(100%) => Algorithm Selection(100%)

2. Current Status

Our group have completed some tasks as mentioned before. We have finished them all on time. Now that the data is cleaned and algorithm has been selected. Currently, we are focusing on the model creation and also Xiaoningsi is working on the data visualization which can show the characteristic of the data in a more vivid way.

3. Project Detail

Raw Data remained like a collections of different categories of attributes and in different values. It is in **autism_screening.csv**

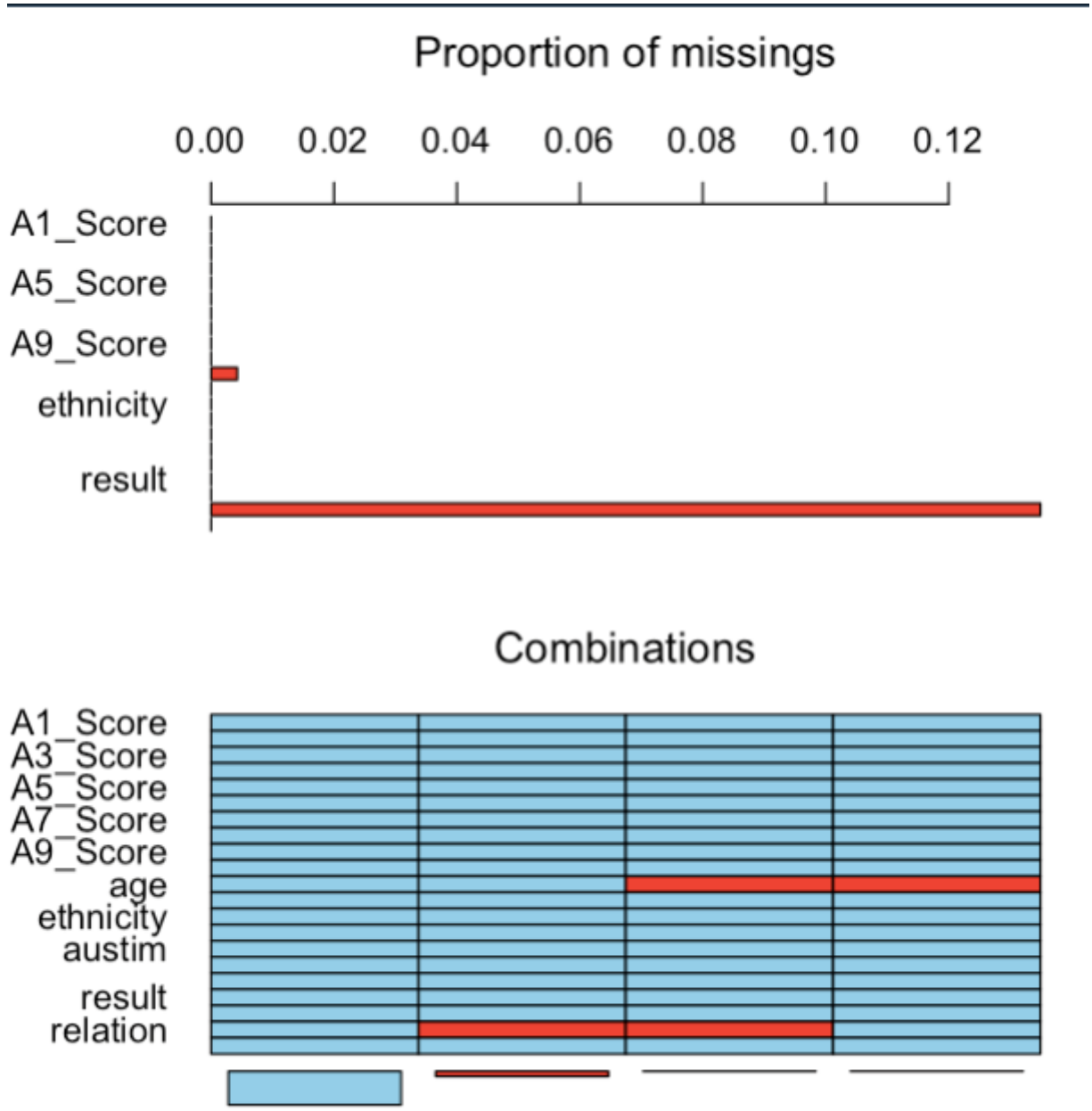
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	age	gender	ethnicity	jundice	austim	contry_of_r	used_app_	result	age_desc	relation	Class/ASD
2	1	1	1	1	0	0	1	1	0	0	26	f	White-Eur	no	no	United Stat	no	6	18 and mo	Self	NO
3	1	1	0	1	0	0	0	1	0	1	24	m	Latino	no	yes	Brazil	no	5	18 and mo	Self	NO
4	1	1	0	1	1	0	1	1	1	1	27	m	Latino	yes	yes	Spain	no	8	18 and mo	Parent	YES
5	1	1	0	1	0	0	1	1	0	1	35	f	White-Eur	no	yes	United Stat	no	6	18 and mo	Self	NO
6	1	0	0	0	0	0	0	1	0	0	40	f	?	no	no	Egypt	no	2	18 and mo	?	NO
7	1	1	1	1	1	0	1	1	1	1	36	m	Others	yes	no	United Stat	no	9	18 and mo	Self	YES
8	0	1	0	0	0	0	0	1	0	0	17	f	Black	no	no	United Stat	no	2	18 and mo	Self	NO
9	1	1	1	1	0	0	0	0	1	0	64	m	White-Eur	no	no	New Zeala	no	5	18 and mo	Parent	NO
10	1	1	0	0	1	0	0	1	1	1	29	m	White-Eur	no	no	United Stat	no	6	18 and mo	Self	NO
11	1	1	1	1	0	1	1	1	1	0	17	m	Asian	yes	yes	Bahamas	no	8	18 and mo	Health car	YES
12	1	1	1	1	1	1	1	1	1	1	33	m	White-Eur	no	no	United Stat	no	10	18 and mo	Relative	YES
13	0	1	0	1	1	1	1	0	0	1	18	f	Middle Eas	no	no	Burundi	no	6	18 and mo	Parent	NO
14	0	1	1	1	1	1	0	0	1	0	17	f	?	no	no	Bahamas	no	6	18 and mo	?	NO
15	1	0	0	0	0	0	1	1	0	1	17	m	?	no	no	Austria	no	4	18 and mo	?	NO
16	1	0	0	0	0	0	1	1	0	1	17	f	?	no	no	Argentina	no	4	18 and mo	?	NO
17	1	1	0	1	1	0	0	1	0	1	18	m	Middle Eas	no	yes	New Zeala	no	6	18 and mo	Parent	NO
18	1	0	0	0	0	0	1	1	1	1	31	m	Middle Eas	no	no	Jordan	no	5	18 and mo	Self	NO
19	0	0	0	0	0	0	0	1	0	1	30	m	White-Eur	no	no	Ireland	no	2	18 and mo	Self	NO
20	0	0	1	0	1	1	0	0	0	0	35	f	Middle Eas	no	yes	United Ara	no	3	18 and mo	Self	NO
21	0	0	0	0	0	0	1	1	0	1	34	m	?	yes	no	United Ara	no	3	18 and mo	?	NO
22	0	1	1	1	0	0	0	0	0	0	38	m	?	no	no	United Ara	no	3	18 and mo	?	NO
23	0	0	0	0	0	0	0	0	0	0	27	f	Black	no	no	United Ara	no	0	18 and mo	Self	NO
24	0	0	0	1	0	0	1	1	1	1	27	m	Middle Eas	no	no	Afghanista	no	5	18 and mo	Self	NO
25	0	0	0	0	0	0	0	1	0	1	42	m	Middle Eas	yes	no	United Ara	no	2	18 and mo	Relative	NO
26	1	1	1	1	0	0	0	1	0	0	43	m	?	no	no	Lebanon	no	5	18 and mo	?	NO
27	0	1	1	0	0	0	0	1	0	0	24	f	?	yes	no	Afghanista	no	3	18 and mo	?	NO
28	0	0	0	0	0	0	0	1	0	0	40	m	Pasifika	yes	yes	United Ara	no	1	18 and mo	Self	NO
29	0	0	0	0	0	0	0	1	0	0	40	m	Middle Eas	yes	yes	Afghanista	no	1	18 and mo	Parent	NO
30	0	0	0	0	0	0	0	1	0	0	48	m	Black	no	no	New Zeala	no	1	18 and mo	Self	NO
31	0	1	1	0	0	0	0	0	1	1	31	m	Middle Eas	no	no	United Kin	no	4	18 and mo	Self	NO
32	0	0	0	0	0	0	0	0	0	0	18	m	White-Eur	no	no	United Kin	no	0	18 and mo	Self	NO
33	1	0	0	1	1	1	1	1	0	1	37	f	White-Eur	no	yes	United Stat	no	7	18 and mo	Self	YES
34	1	1	0	0	0	0	1	0	0	1	55	f	Others	no	no	New Zeala	no	4	18 and mo	Self	NO
35	1	1	1	1	1	1	1	1	1	1	18	f	White-Eur	yes	no	South Afric	no	10	18 and mo	Self	YES

Firstly, we have to manage the exceptional values in the csv file.

Secondly, we combine about multiple attributes into one attribute, from which we guarantee a better data analysis in the future.

```r
aurism <- read.csv("/Users/argebell/Desktop/研一课程/programming/group\ project/autism_screening.csv")#导入数据集
summary(aurism)#查看数据集内部结构
###异常值处理###
aurism$age[aurism$age==383] <- NA#将 383 视为缺失值
summary(aurism$age)#查看处理后 age 的情况
###属性值合并###
aurism$contry_of_res = as.factor(ifelse(aurism$contry_of_res %in% c("Canada", "Mexico", "Nicaragua", "United States"), "North America", ifelse(aurism$contry_of_res %i
aurism$ethnicity[aurism$ethnicity %in% c("?","others", "Hispanic", "Latino", "Pasifika", "Turkish")]="Others"#将 ethnicity 的部分属性值合并
aurism$relation = as.factor(ifelse(aurism$relation =="Self", "Self", ifelse(aurism$relation =="?", "?", "Other")))#将 relation 划分为 Self 和 Other
aurism$age=as.factor(ifelse(aurism$age <= 27 , "≤27", ">27"))#将 age 划分为"≤27"和"> 27"
summary(aurism)#查看数据集内部结构
###查看缺失值分布###
library(VIM)#导入 vim 包
aurism[aurism == "?"] <- NA #将?转化为 NA
aggr(aurism,prop=T,numbers=T)#缺失值分布可视化
pMiss <- function(x){sum(is.na(x)) / length(x) * 100}#定义查看缺失数据比例的函数
apply(aurism, 2, pMiss)#查看各属性缺失值比例
###利用多重插补法处理缺失值###
library(lattice)#导入 lattice 包
library(MASS)#导入 MASS 包
library(nnet)#导入 nnet 包
library(mice)#导入 mice 包
miceMod <- mice(aurism[, !names(aurism) %in% "Class"], m = 5, seed = 1234)
aurism1 <- complete(miceMod, action = 3) # 生成完整数据
anyNA(aurism1)#查看插补后的数据集时候存在缺失值
aurism$age = aurism1$age#将插补后的 age 放入 aurism 中
aurism$relation = aurism1$relation#将插补后的 relation 放入 aurism 中
#进行 mice 插
summary(aurism)#查看数据集内部结构
###AVF 检测离群值###
freq_matrix <- table( unlist( unname(aurism) ) )
aurism[,"Score"] <- apply(aurism,1,function(x) { sum(freq_matrix[x]) / length(x) })#计算 AVF
aurism1 <- aurism[order(aurism$Score),]#将数据集按照 Score 排列
summary(aurism1)#查看数据集内部结构
aurism2<-aurism1[36:704,1:22]#删除 AVF 值较小的 5%的记录 summary(aurism2)#查看数据集内部结构
###保存预处理后的数据集###
write.csv(aurism2, "/Users/argebell/Desktop/aurism_processing.csv")#将处理后的训练集导出
```

In addition, we can see from the picture below that there are lots of attributes remaining the Null values. So we should manage with none values and do some filling.



What is more, outlier handling is a must, and you can see it in the R code.

After we introduced different methods of data processing to this raw data, we have got plenty of results.

Firstly, we have turned the location string into certain numbers. Then after that, we drew a correlation matrix.

Then, with the results of data normalization, we then turn them into 2 sets which are training set and test set. And the results of data processing is just in **autism_processing2.csv**

In terms of variable selection, we use Lasso Variable selection, Optimal Subset selection, and Random Forest to select adult autism and the main influencing factors.

In this section, we choose to use Lasso regression for variable selection and introduce L1 norm penalty term. L1 Norm Table. The expression is as follows $L_1 = \lambda \sum_{j=1}^{p} |\beta_j|$

Select adjustment parameters through cross validation λ , and select from 10^(-3) to 10^(10) for λ Values, calculating each λ of Cross validation error, and then output the minor-error λ Value, at this time λ is About 0.002971. The cross validation error value is 0.068835, and then the corresponding non-zero variable coefficient is output. It is found that no variable is eliminated, so all variables will be the variables to be considered. All will be as the main influencing factor of adult autism selected by Lasso variable selection method.

4. Next Step

Recently, we are focusing on the model creation and the data visualization which are both crucial. And our core models are now under construction. More working materials have been stored in wechat groups.