Modern Data Mining - Final Project

Predicting Dementia Among Elderly People

Jia Xu Yuqin Zhang Zejia Cai

I. Abstract

Alzheimer's disease (AD) is the most common type of dementia among elderly people that leads to memory loss. More severely, it can affect the patient to carry out daily activities. AD is a progressive disease and usually starts slowly, but changes in the brain can begin many years before the appearance of first symptoms. Age has also shown to be associated with the risk of developing AD.

Our data consists of Magnetic Resonance Imaging information for demented and nondemented elderly adults. In this study, we present several ways of building classifiers to predict whether a subject will be diagnosed to develop dementia.

The data was released by Open Access Series of Imaging Studies (OASIS).

II. Description of the Data

The dataset is a longitudinal collection of MRI scan history of 150 elder adults aged between 60 and 98. Subjects may be scanned more than once, and there are 373 imaging sessions recorded in total. The following table showcases the features that the original data contains:

Variable Name	Description
Subject.ID	The unique identification of each subject
MRI.ID	The unique identification of each scan session
M.F	Gender of the subject
Age	Age of the subject
Hand	Dominant Hand
EDUC	Years of education
SES	Socialeconomic status
MMSE	Mini Mental State Examination Score
eTIV	Estimated total intracranial Volume
nWBV	Normalized whole brain volume
ASF	Atlas scaling factor
MR.Delay	MR Delay Time
CDR	Clinical Dementia Rating

Here are some more detailed explanations of the terms mentioned above:

Mini Mental State Examination (MMSE): This is a 30-point questionnaire which has been widely adopted to measure cognitive functions among elderly people.

Estimated total intracranial volume (eTIV): This is an estimated value of the maximum pre-morbid brain volume.

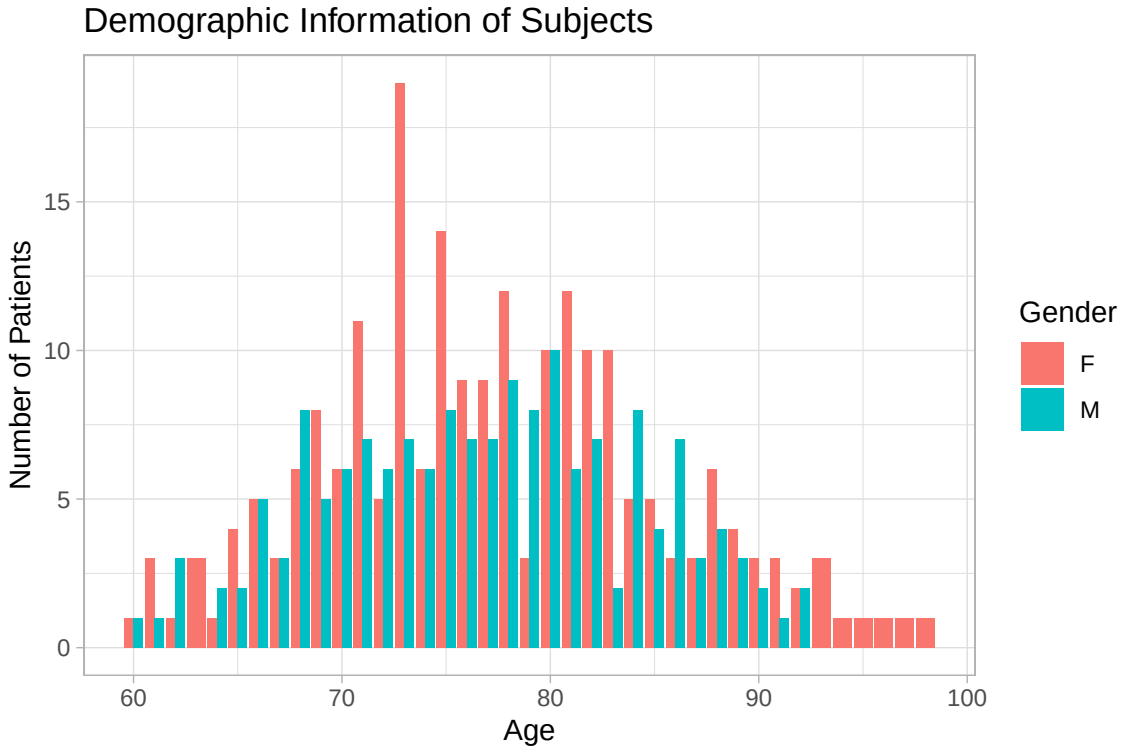
Atlas scaling factor (ASF): This is a volume-scaling factor that standardizes the head size based on differences in human anatomy.

MR Delay: A delayed MR is performed a few minutes after the injection of the contrast agent. The delayed contrast enhancement might reveal different biological information.

Clinical Dementia Rating: This is a globally accepted measure of the overall severity of dementia.

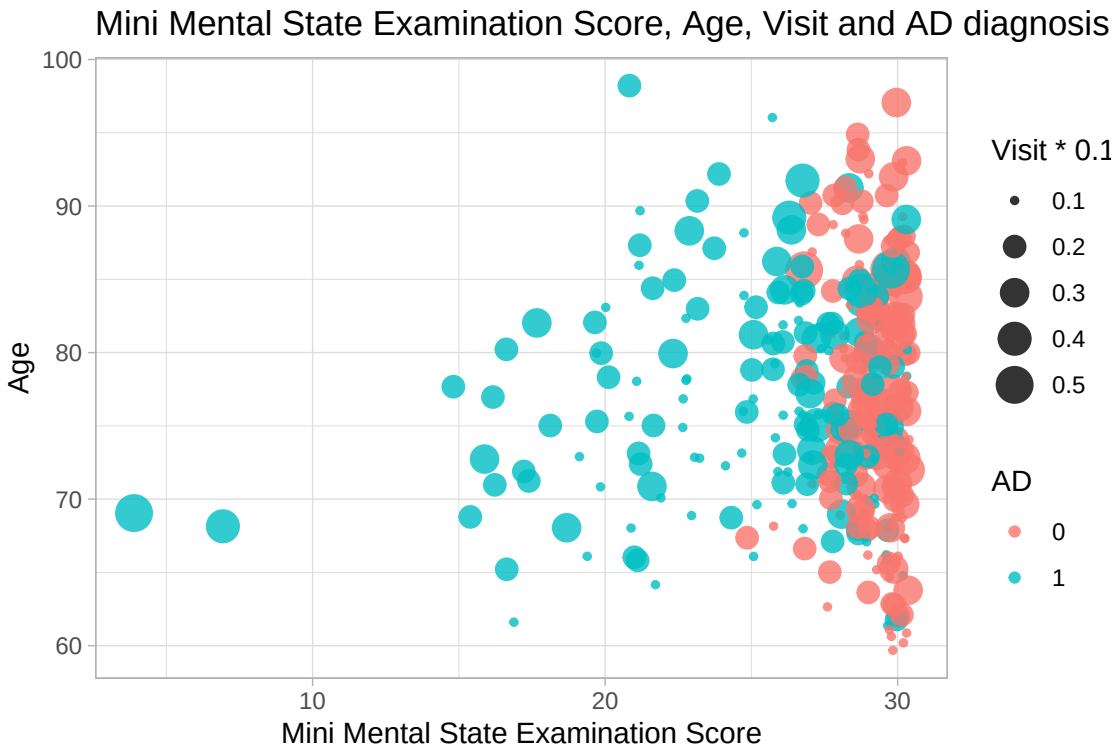
III. Data Cleaning and Preparation

IV.Exploratory Data Analysis



TODO:Change size legend! scale_size_manual doesn't work??

AD is associated with lower mini mental state examination score. Age does not pose a significant influence on the examination score and the diagnosis of AD. Two subjects are having significantly lower mini mental state examination score. They are also visiting the hospital very often.



##TODO: more visualizations?

Feature engineering maybe?

V. Model Building

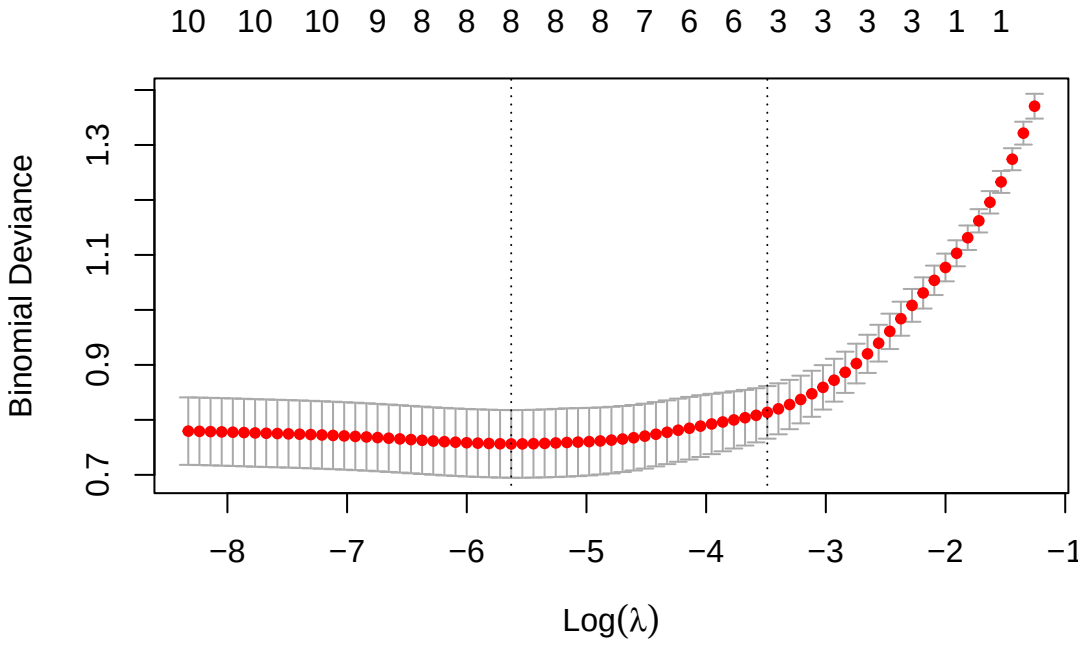
Data Splitting

We split the data into three sets: training, testing and validation. The training set will be used to fit a model; the testing set will be used report a model's effectiveness; and the validation set will be held until the end to evaluate our final model.

Model 1: LASSO Logistic Regression Model

- Model Fit

We first fit a logistic regression model. We selected a sparse model by using LASSO regularization technique. The criteria is set to be deviance and 10 fold Cross Validation is applied.



We first choose the set of variables which give the smallest deviance. The variables selected are: Gender, Age, EDUC, SES, MMSE, eTIV, nWBV, ASF.

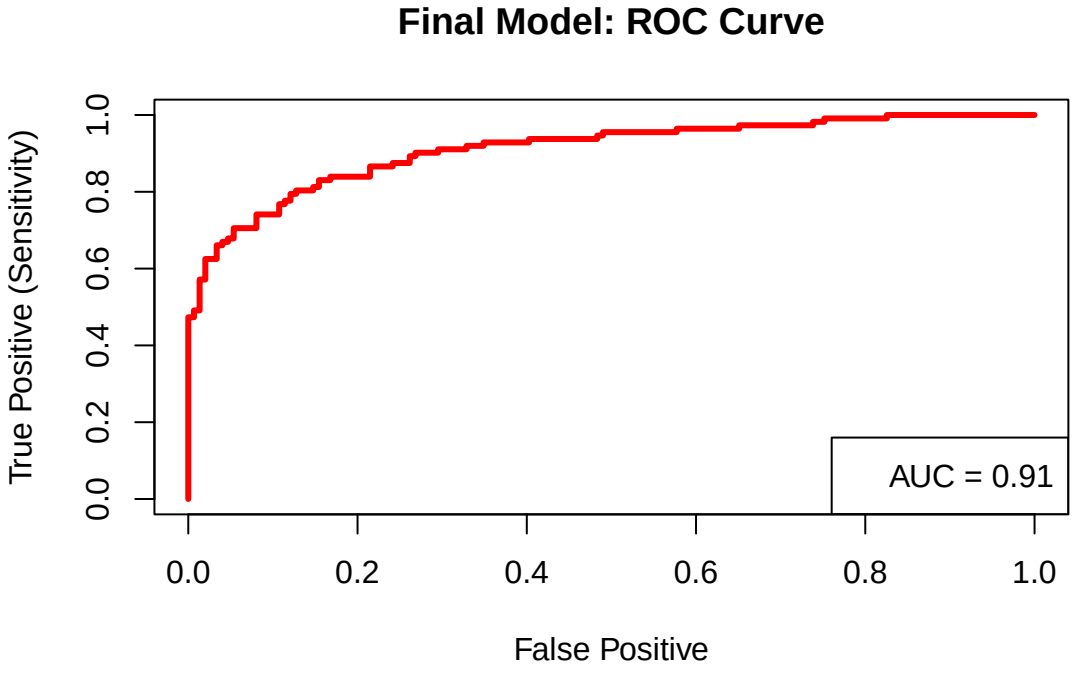
- Fine Tuning

We refit the model using backward selection to drop any variables with significance level $p < 0.05$. In the end, four variables remained, which are Gender, Age, MMSE, nWBV.

- Analysis

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



Here, we simply assume that it costs equally to mislabel a subject to be AD as it does to mislabel a non-AD.

Thus, we will choose the threshold to be 0.5.

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	22	4
$\hat{Y} = 1$	6	23

xxx

Random Forest

xxx

Final Model:

Model Evaluation

Conclusion

Reference

Fulton, L.V.; Dolezel, D.; Harrop, J.; Yan, Y.; Fulton, C.P. Classification of Alzheimer's Disease with and without Imagery Using Gradient Boosted Machines and ResNet-50. Brain Sci. 2019, 9, 212. https://doi.org/10.3390/brainsci9090212