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. The score has the following 5 values:

0 - Normal

0.5 - Very Mild Dementia

1 - Mild Dementia

2 - Moderate Demantia

3 - Severe Demantia

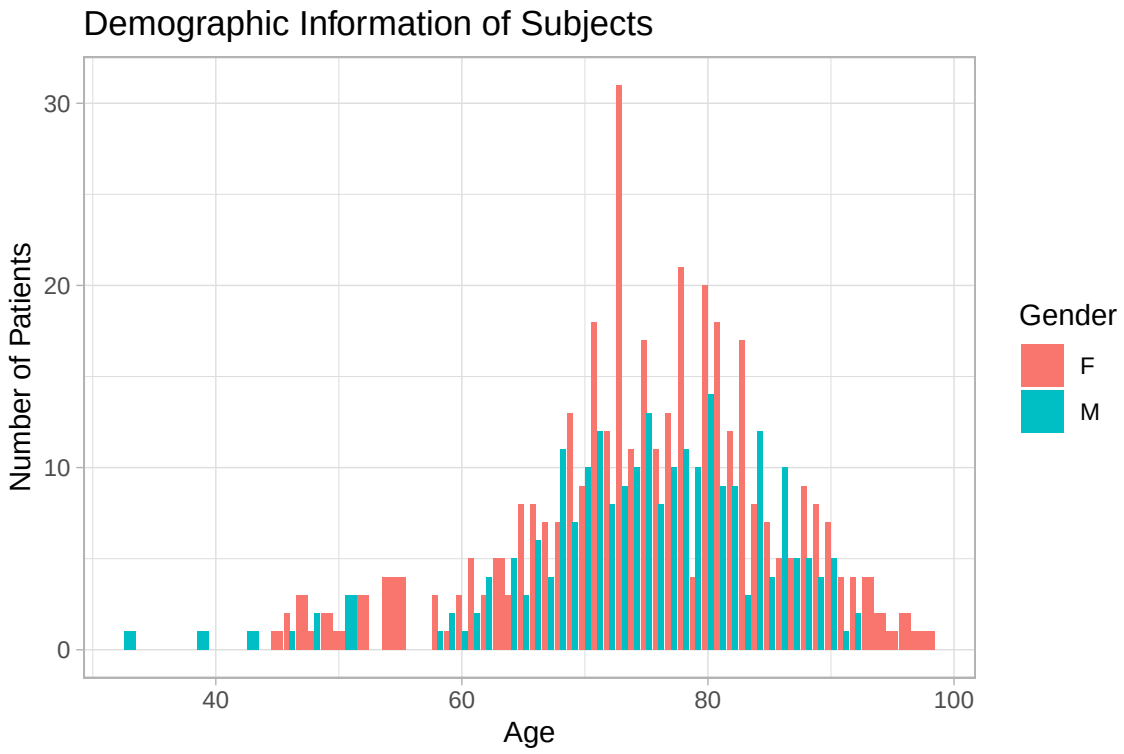
Target Value

We will predict whether a patient has developed dementia based on the Clinical Dementia Rating. If a patient has CDR score = 0, the subject has no AD, thus receiving a label 0. If the score > 0, the subject has developed AD, thus will be marked with 1.

III. Data Cleaning and Preparation

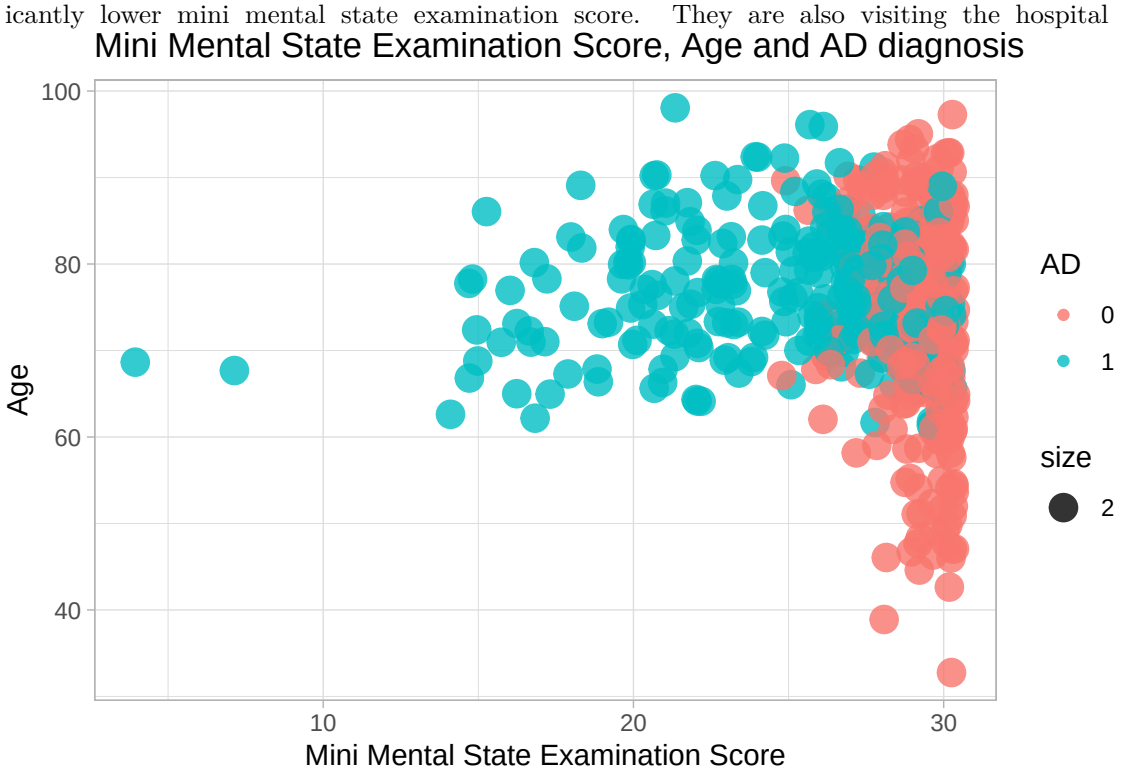
We combine the cross sectional dataset with the longitudinal dataset.

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.



V. Model Building

Data Splitting

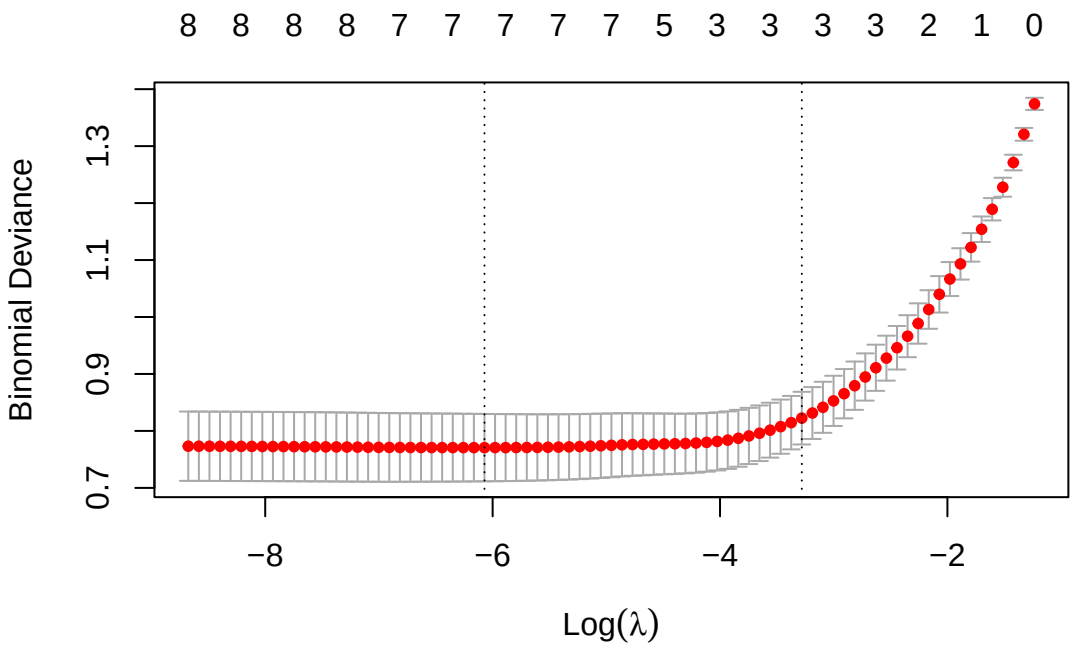
We split the data into three sets: training, testing and validation. The training set is used to fit a model; the testing set is used to report a model's effectiveness; the validation set is held until the end to evaluate our final model. We randomly select 70% of the data to be the training set, 15% of the observations to be the testing set and the remaining 15% to the the validation set.

Model 1: Logistic Regression

Model Fitting

We first fit a logistic regression model. We select a sparse model by using LASSO regularization technique and use the **cv.glmnet()** function to implement cross validation. The criteria is set to be deviance and 10-fold cross validation is applied.

The plot below shows how the deviance varies with λ and the number of non-zero coefficients.



We start with choosing the set of variables which give the smallest cross-validated error. The 7 variables selected are: Gender, Age, EDUC, SES, MMSE, eTIV and nWBV.

Fine Tuning

Fitting the logistic model using these variables, we notice that not all the variables are significant at level 0.05. Thus, based on this model, we proceed to perform backward selection until all the remaining variables are significant at level 0.05.

In the end, only three variables remain, which are Gender, MMSE and nWBV. In fact, we notice that this is the same set of variable corresponding to `lambda.1se`, which is the largest value of λ such that the cross-validated error is within 1 standard error of the minimum cross-validated error.

Our final logistic model gives the following result:

```
Call:
glm(formula = AD ~ Gender + MMSE + nWBV, family = binomi

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9871  -0.6112  -0.2878   0.3708   2.4286

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  36.3832     4.2921   8.477  < 2e-16 ***
GenderM       0.8221     0.2912   2.824 0.004749 **
MMSE         -0.9462     0.1094  -8.652  < 2e-16 ***
nWBV        -14.1105     3.7663  -3.746 0.000179 ***
```

Analysis

Based on the summary table of the model, the logit function is given by:

$$\begin{aligned} logit(P(AD = 1 \mid Gender, MMSE, nWBV)) &= log(\frac{P(AD = 1 \mid Gender, MMSE, nWBV)}{P(AD = 0 \mid Gender, MMSE, nWBV)}) \\ &= 36.38 + 0.82 \cdot Gender(Male) - 0.95 \cdot MMSE - 14.11 \cdot nWBV \end{aligned}$$

where

$$P(AD = 1 \mid Gender, MMSE, nWBV) = \frac{exp(36.38 + 0.82 \cdot Gender(Male) - 0.95 \cdot MMSE - 14.11 \cdot nWBV)}{1 + exp(36.38 + 0.82 \cdot Gender(Male) - 0.95 \cdot MMSE - 14.11 \cdot nWBV)}$$

Here, we will assume that it costs equally to mislabel a subject to be AD as it does to mislabel a non-AD. Thus, we will set the threshold to be 0.5. That is,

$$\hat{AD} = 1 \text{ if } \hat{P}(AD = 1 \mid Gender, MMSE, nWBV) > 0.5$$

Using the testing dataset to evaluate the performance of this model, we see that the misclassification error rate is 0.198. The confusion matrix is given below.

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	41	11
$\hat{Y} = 1$	7	32

This model has a sensitivity rate of 0.744 and a specificity rate of 0.854.

Findings TODO

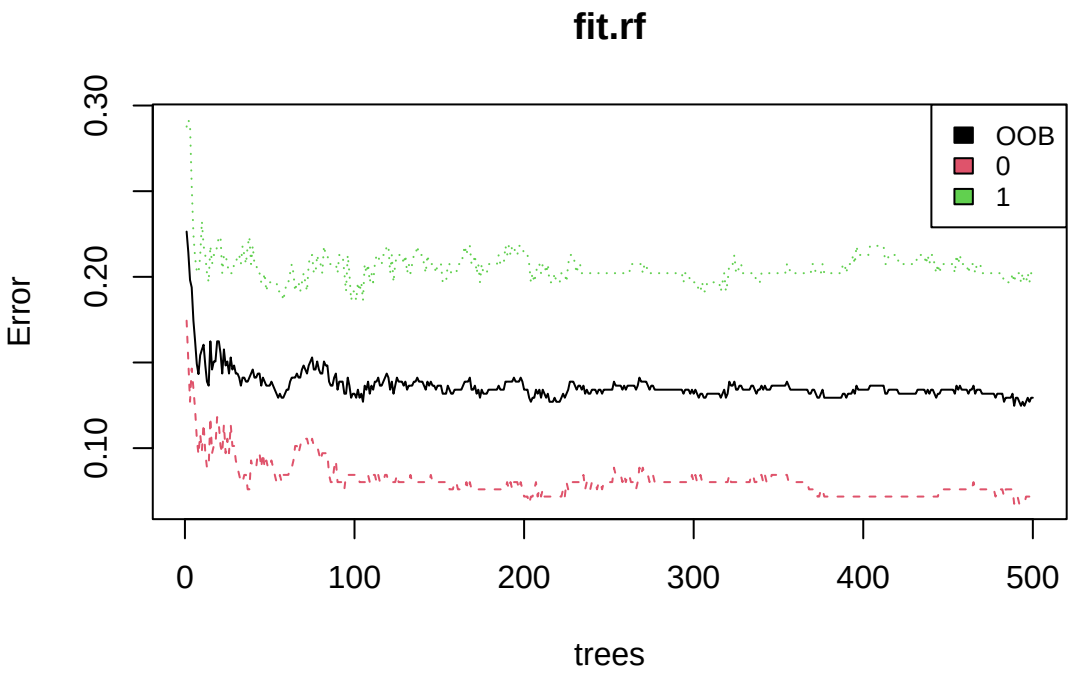
With only 3 variables left in the logistic regression model, the result is rather simple to interpret. The most important variable is...

Model 2 Random Forest

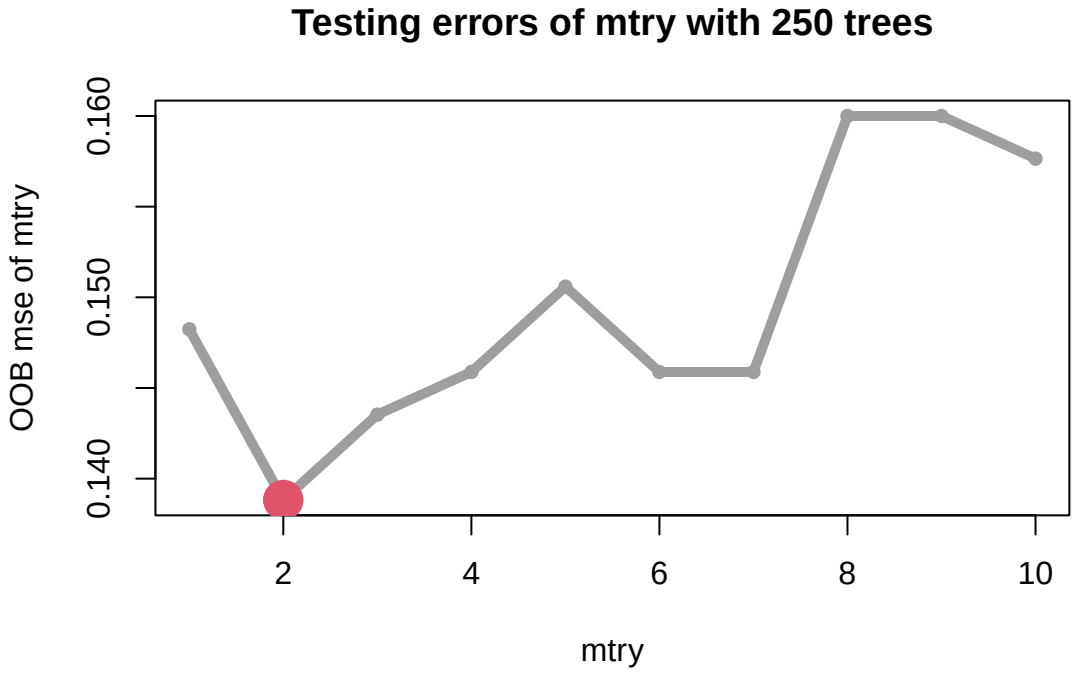
Model Fitting and Fine Tuning

The second model that we build is a random forest model. We first set `mtry` (number of randomly chosen predictors at each split) to be 3, which is the square root of the number of predictors. We set `ntree` to be 500. The split criterion is set to be misclassification error.

By plotting the error rate v.s number of trees, we decide to use 300 trees in order to settle down the OOB testing errors.



Then, by setting `ntree=300`, we want to compare effects of different `mtry`. Thus, we loop `mtry` from 1 to 10 and return the testing OOB errors for each of the model. In the end, we decide to use `mtry=2` which gives the minimum error rate.



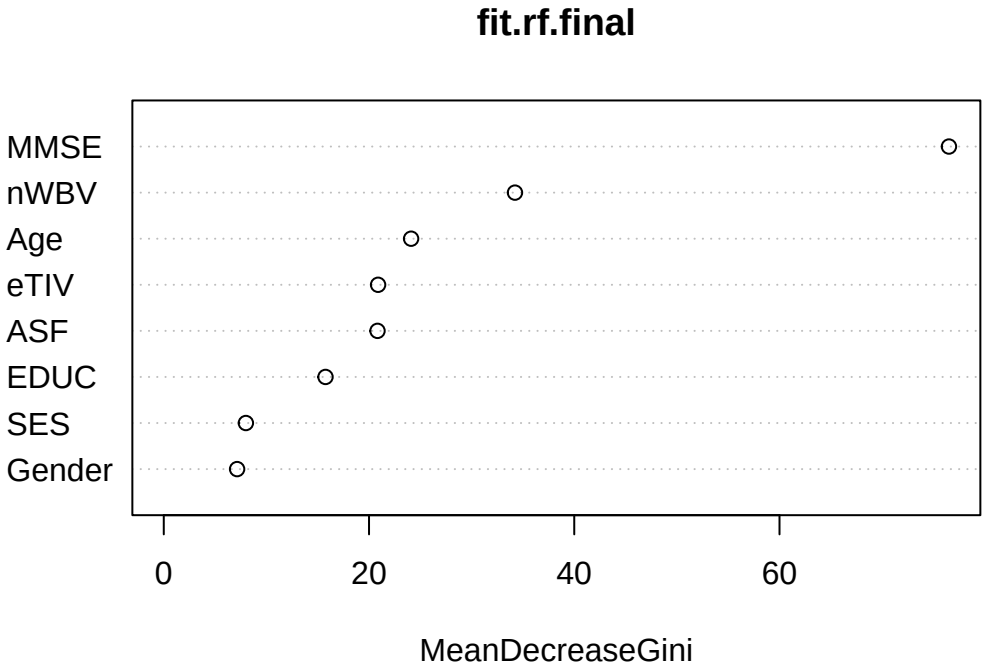
Analysis

Using 0.5 as threshold to determine the class of instances in the test dataset, the misclassification rate of our final random forest model is 0.143, the sensitivity rate is 0.814 and the specificity rate is 0.896. The confusion matrix is also shown below and we can see that the model roughly achieves a balance to predict positive and negative class.

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	43	8
$\hat{Y} = 1$	5	35

Findings

From the variable importance plot of our final random forest model, we can see that MMSE is the most important feature in this model.



The plot below demonstrates the distribution of minimal depth among all the trees in the forest. The vertical bar is the mean of minimal depth for each feature. This plot can give us a clearer idea of the role that each feature plays in our model.

We can see that MMSE and nWBV are more likely to be the root of the tree compared to other variables. The average minimal depths of MMSE and nWBV are around 1.5, suggesting that many dementia observations can be separated effectively on the basis of these two variables.

Model 3 & 4: Boosting

Model Fitting and Fine Tuning

For boosting, we implement both gradient boosting and extreme gradient boosting model. Both models turn out to give similar result.

We use the grid search method which iterates over many possible combinations of hyperparameter values, thus determining the best-performing set of hyperparameters.

In the case of gradient boosting, the `distribution` is set to be `multinomial`. We then look at 135 models with various combinations of learning rate, tree numbers, tree depth and minimum number of observations in the end node. We apply 10-fold cross validation and record the minimum cross validated error for each of the model.

In the case of extreme gradient boosting, we apply 10-fold cross validation and set the `early_stopping round` to be 50 to avoid overfitting. This means that the algorithm will be forced to stop if we do not see an

improvement of the model performance in 50 iterations. Similar to the implementation of gradient boosting, we use grid search and test 240 models in total. We tune hyperparameters include learning rate, tree depth, minimum loss reduction for a split and penalty on the number of leaves in a tree,

Analysis

The final gradient boosting model gives an error rate of 0.187. The final extreme gradient boosting model gives an error rate of 0.165.

- Predict

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	11	1
$\hat{Y} = 1$	37	42

Findings

Final Model

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

Knight Alzheimer Disease Research Center, CDR Scoring Table: https://knightadrc.wustl.edu/professionals-clinicians/cdr-dementia-staging-instrument/cdr-scoring-table/