Modern Data Mining - Final Project

Alzheimer's Disease Classification

Jia Xu Yuqin Zhang Zejia Cai

I. Abstract

Alzheimer's disease (AD) is the most common type of dementia which leads to memory loss and decline in thinking. AD is a progressive disease and usually starts slowly, but changes in the brain can begin many years before the appearance of first symptoms.

In this study, we aim to use Magnetic Resonance Imaging (MRI) data for both demented and nondemented adults to build classifiers that predict whether a subject will be diagnosed to develop dementia. Two datasets are used: one deals with cross-sectional MRI data for adults aged between 18 to 96, and the other deals with longitudinal MRI data for older adults between 60 to 96.

Our datasets can be found in Kaggle. The data was originally collected and released by the Alzheimer's Disease Research Center at Washington University and Open Access Series of Imaging Studies (OASIS).

II. Description of the Data

After data cleansing, we append the longitudinal and cross-sectional data and obtain 608 valid observations in total.

The following table showcases the features that the data contains:

Variable Name	Description
ID	The unique identification of an MRI 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	MRI 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 of individuals, especially 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 the subject has developed dementia based on the Clinical Dementia Rating. If the subject has a zero CDR score, the subject has no AD, thus receiving a label of 0. If the score is greater than zero, the subject has developed AD, thus will be marked with 1.

III. Data Cleaning and Preparation

We append the cross sectional dataset and the longitudinal dataset in order to get as many MRI session records as possible. In order to do it, we only keep the common variables of the two datasets and rename the variables so that they match across the datasets.

Null values exist in 4 columns: `EDUC` (Years of Education), `SES` (Social-Economic Status), `MMSE` (Mini Mental Score Exam Score), `CDR` (Clinical Dementia Rating). Since the target variable is produced according to CDR, we drop all the observations with no CDR score. For the categorical variable `SES`, we impute missing values with the mode. For the numerical variables `EDUC` and `MMSE`, we impute the missing values with the mean.

We further drop the column `Hand` because all subjects are reported to be right-handed.

In the end, we obtain 608 valid observations with the following variables: `Gender`, `Age`, `EDUC`, `SES`, `MMSE`, `eTIV`, `nWBV`, `ASF`, `CDR`. We create our target variable `AD` based on the value of `CDR`. Specifically, we let `AD` $= 0$ for `CDR` $= 0$, and `AD` $= 1$ for `CDR` > 0. There are 341 Non-AD observations and 267 AD observations.

Below, we provide a summary statistic table for the numerical variables in our dataset.

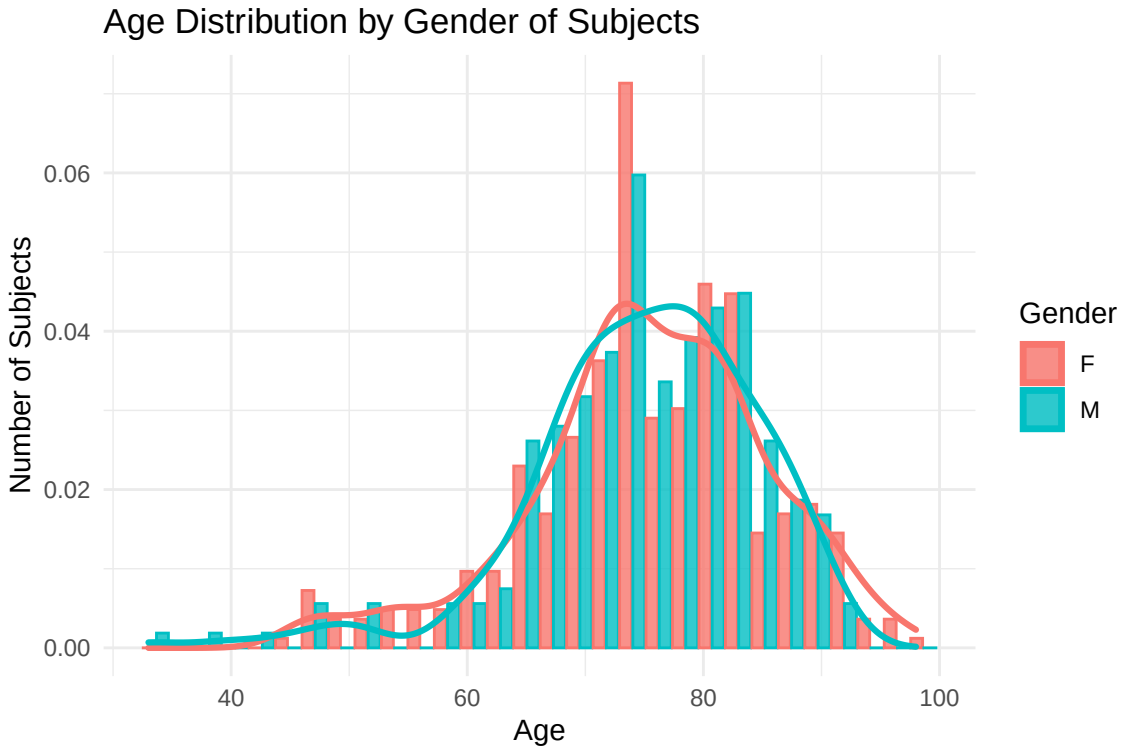
variable	min	max	median	mean	sd
Age	33.000	98.000	76.000	75.209	9.865
ASF	0.876	1.587	1.202	1.204	0.135
EDUC	1.000	23.000	12.000	10.184	6.058
eTIV	1106.000	2004.000	1460.000	1477.062	170.654
MMSE	4.000	30.000	29.000	27.234	3.682
nWBV	0.644	0.847	0.736	0.737	0.043
SES	1.000	5.000	2.000	2.442	1.098

IV.Exploratory Data Analysis

In this section, we create several visualizations to investigate the distribution of data and to see the relationship between different variables.

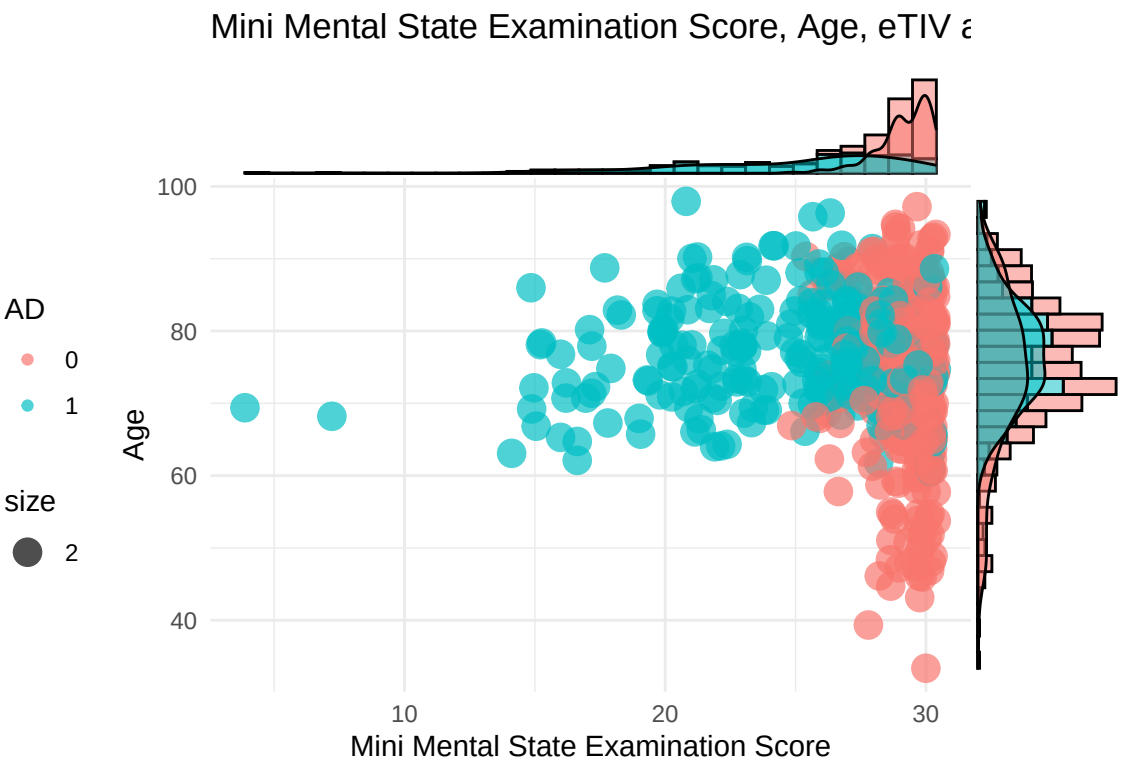
First, we are interested in the demographic distribution. A histogram of age by gender showcases that the distributions of age and gender are roughly balanced. In addition, we notice that the age of subjects tends to be high. Most subjects are aged between 60 to 85 years old. Thus, the data deals mainly with elderly patients.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

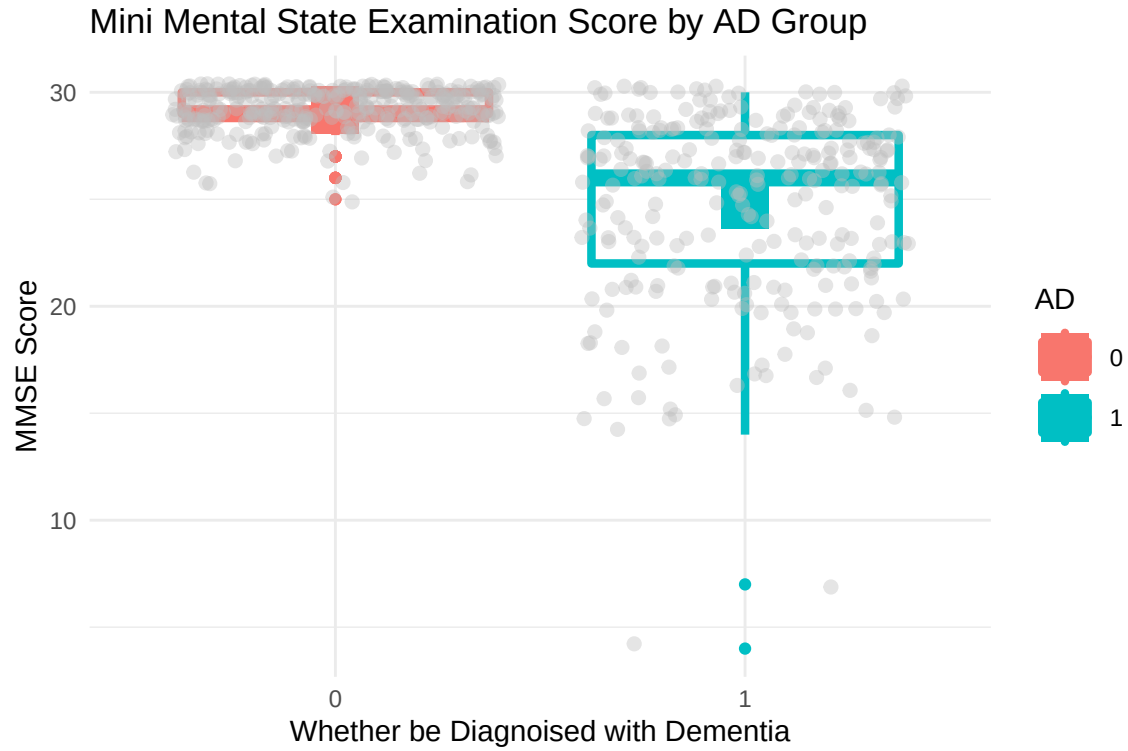


Next, we visualize the relationship between Mini Mental State Exam score, Age and whether a patient is diagnosed with dementia. On the top, the marginal density plot depicts the distribution of MMSE score by AD label; on the right, the marginal density plot is for the distribution of age.

We see that subjects with AD tend to have lower MMSE score, but there is no clear relationship between age and AD in this plot. Age and MMSE score does not have a close association either.

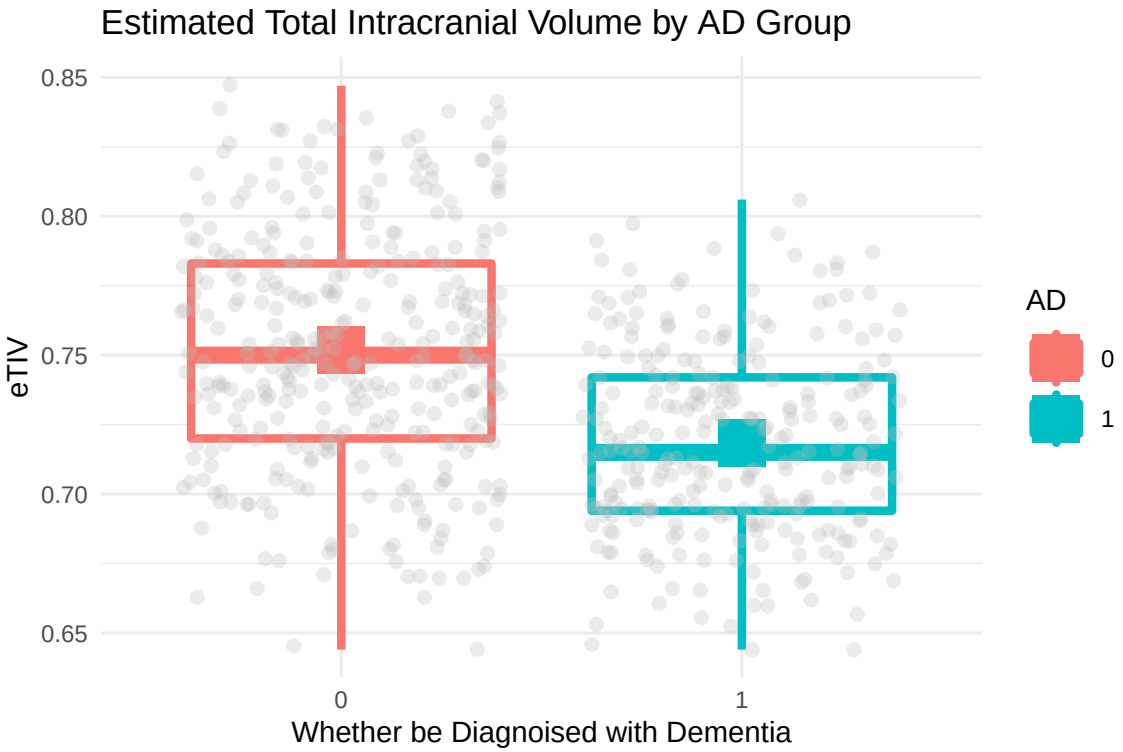


We dive deeper into the relationship between MMSE score and AD diagnosis. Indeed, from the box plot below, we can see that those who are diagnosed with AD have a lower MMSE score in general. The median (the middle horizontal line of the box plot) and mean (marked by the square) are both significantly lower for the AD group.

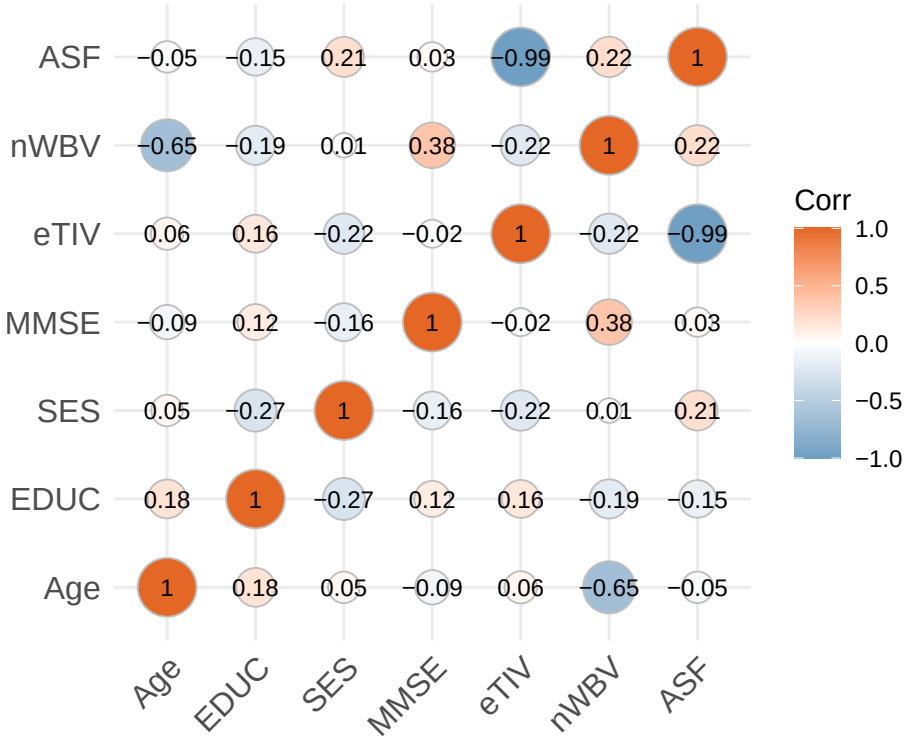


We also notice that the estimated total intracranial volume (eTIV) differ between demantia and non-demantia

group. People with demantia tend to have a lower eTIV value.



Lastly, we plot the correlation between numerical variables of our dataset. ASF and eTIV have very high negative correlation. nWBV and Age also have a relatively high negative correlation. We choose to keep all these variables for the modeling part. In fact, ASF does not turn out to be significant in any of the models, but eTIV does have some significance.



V. Model Building

Data Splitting

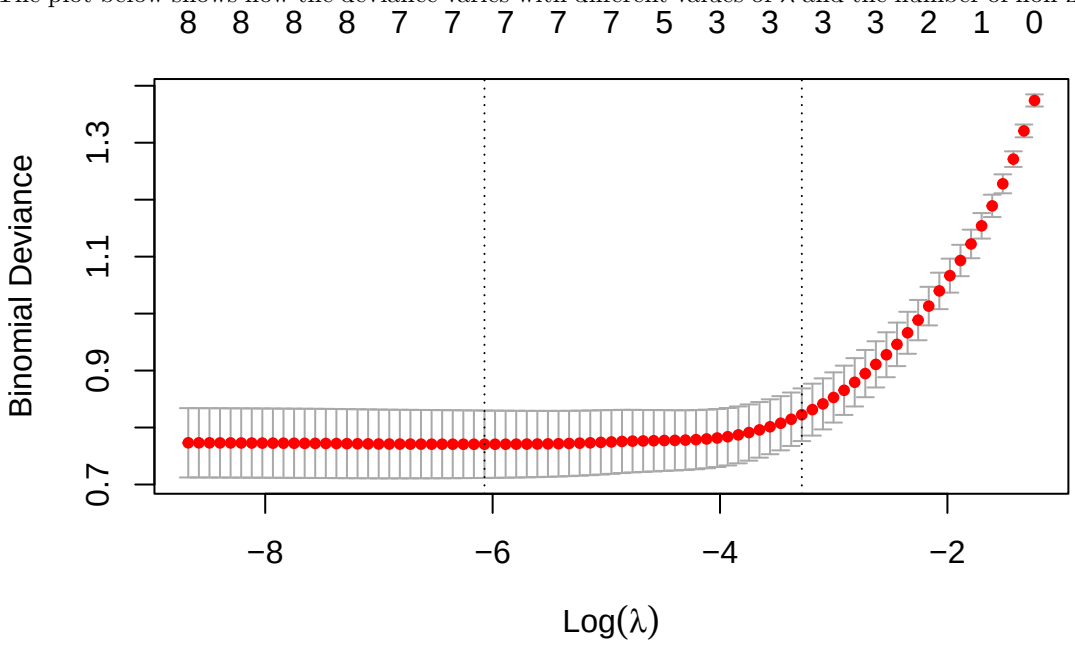
We split the data into three sets: training, testing and validation. The training set is used to fit a model; the testing set is used to report a model's effectiveness; the validation set is held until the end to evaluate our final model. We randomly select 70% of the data to be the training set, 15%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

Model 1: Logistic Classification

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 different values of 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 exactly the same set of variables 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, the misclassification error rate turns out to be 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

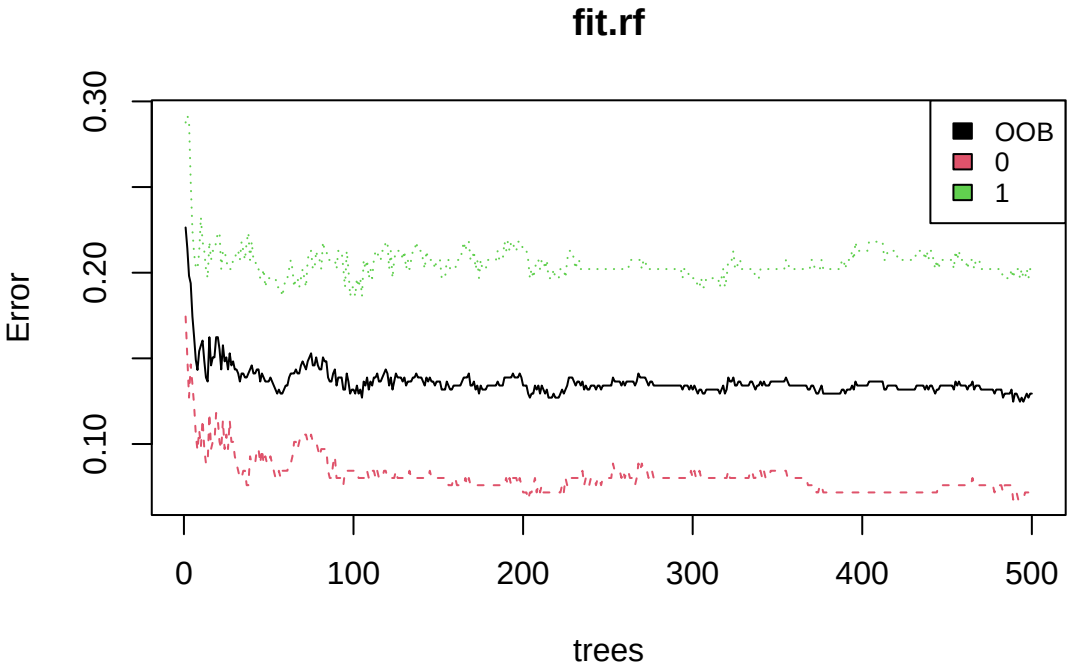
With very few variables remaining in the logistic regression model, the result is rather simple to interpret. The model suggests that MMSE, nWBV and age are the top 3 important variables in predicting one's probability of developing demantia. The log odds is a decreasing function of MMSE and nWBV. Holding other variables constant, when MSSE decreases, the probability of being diagnosed with AD increases. Similarly, decrease in nWBV also implies higher probability of AD diagnosis. On the other hand, the log odds is an increasing function of age. Specifically, when age is increased by 1, the log odds increases 0.82.

Model 2 Random Forest

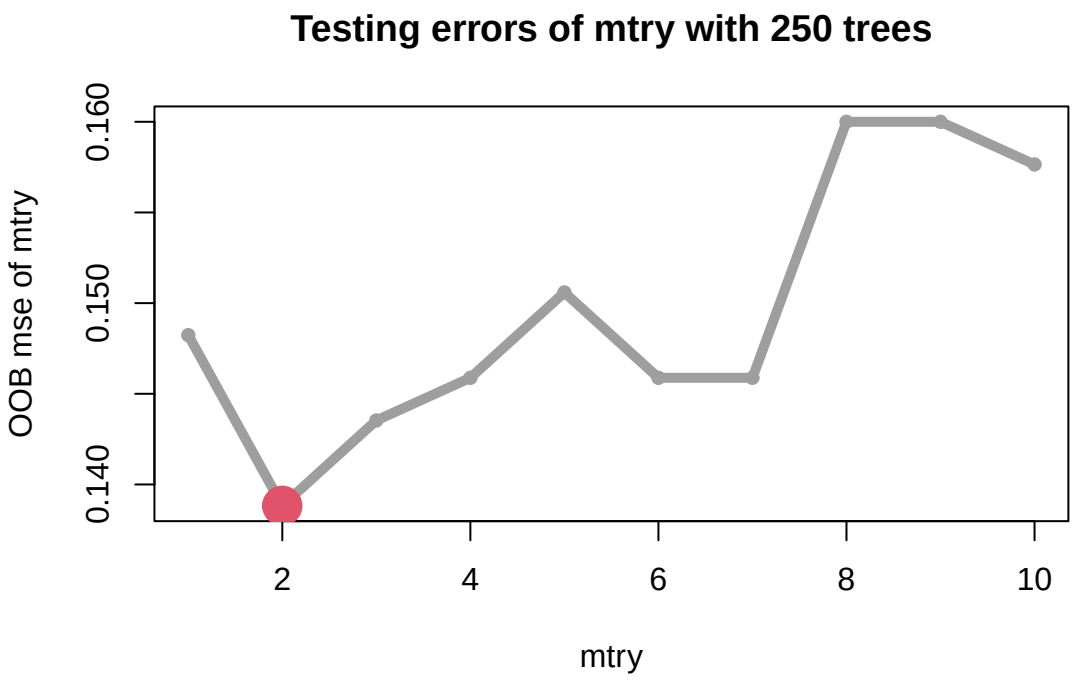
Model Fitting and Fine Tuning

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.



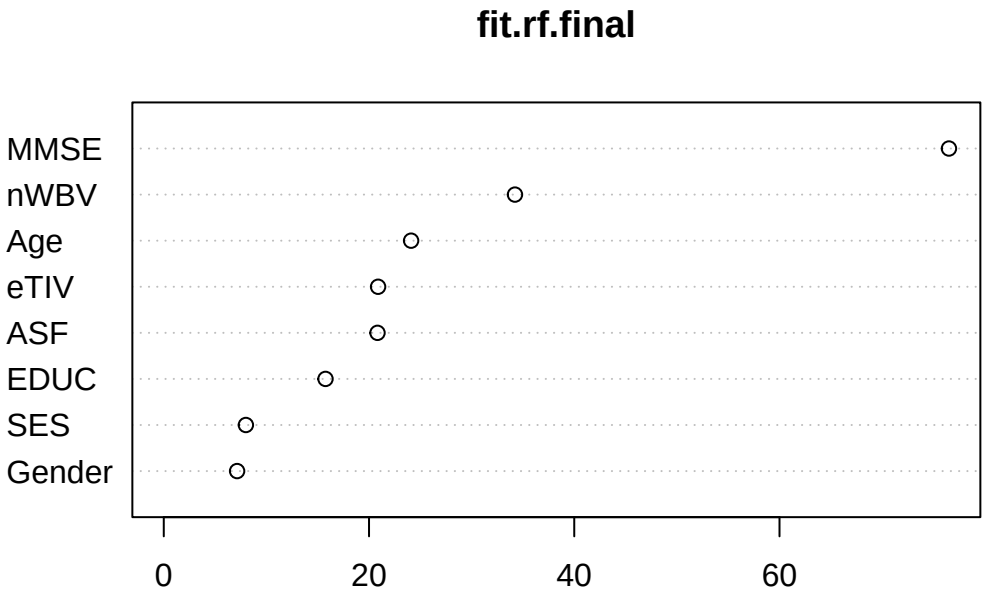
Analysis

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 given 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

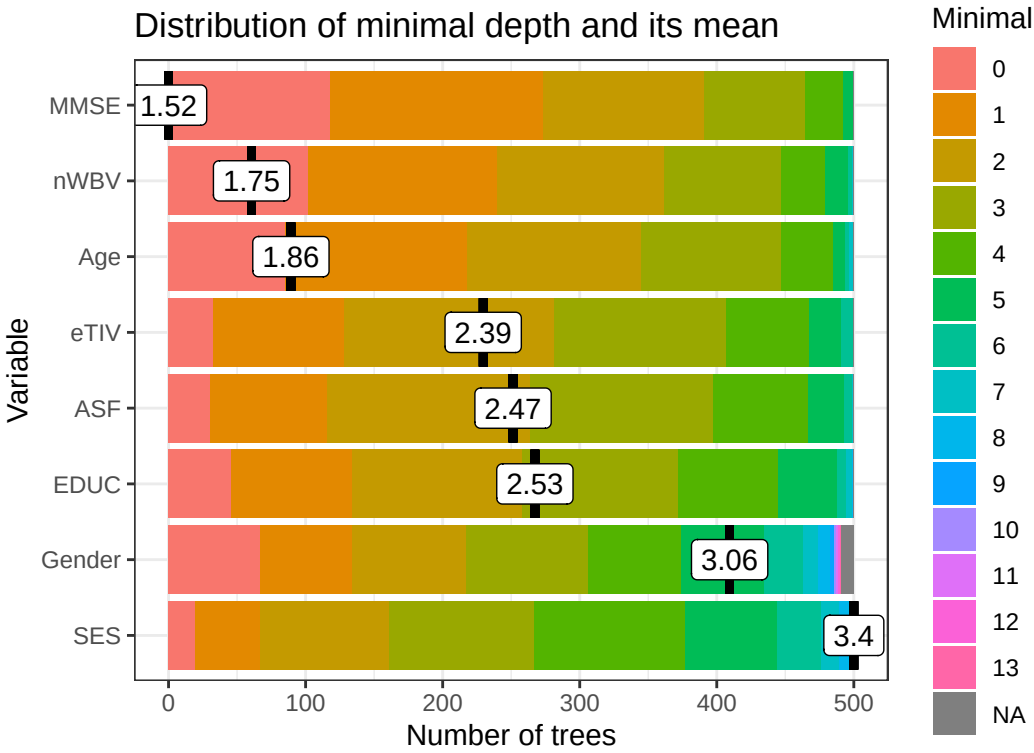
Findings

From the variable importance plot of our final random forest model, we can see that MMSE is the most important feature in this model. nWBV and Age also have high importance. This result corresponds with the finding of logistic classification 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 produce similar results on the testing dataset.

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

Analysis

Both gradient boosting model and extreme gradient boosting model give an error rate of 0.143.

The confusion matrix of gradient boosting model is:

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	44	9
$\hat{Y} = 1$	4	34

The confusion matrix of extreme gradient boosting model is:

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	42	7
$\hat{Y} = 1$	6	36

We see that on this specific testing dataset, gradient boosting model has fewer false discovery cases than the extreme gradient boosting model. On the other hand, extreme gradient boosting model has a better balance between false negative and false positive rate. Overall, the performance of the two boosting methods are rather similar.

Findings

Below, the table shows the top 5 important features in the Extreme Gradient Boosting model.

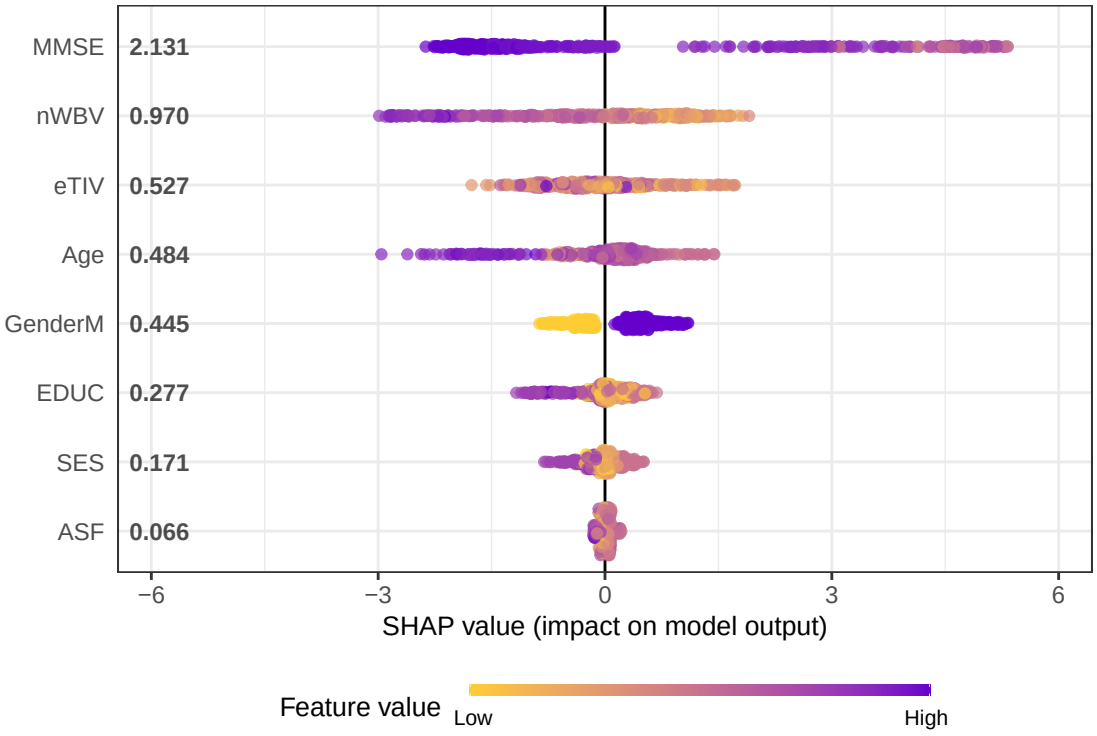
Feature	Gain	Frequency
MMSE	0.4292515	0.1341044
nWBV	0.1784368	0.2400750
eTIV	0.1602309	0.2678962
Age	0.1145990	0.1750547
EDUC	0.0494781	0.0903407

The `gain` measures the contribution of the feature to the model. We can see that MMSE has a significantly higher `gain` than other variables. The `frequency` measures the relative proportion of times for which a feature occurs in the trees. We see that nWBV and eTIV have high `frequency`. MMSE and Age also have relatively high `frequency`.

We now display the first tree in the Extreme Gradient Boosting model. We see that the first cut is based on MMSE and the second cut is based on nWBV. Age and eTIV appear in subsequent cuts. This roughly corresponds to the feature importance table shown above.

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```

We also create a SHAP (SHapley Additive exPlanation) visualization plot for the Extreme Gradient Boosting model. The SHAP value demonstrates the contribution of each feature value to the prediction of dementia. We can see that high values of Mini Mental State exam score, low values of normalized whole brain volume and estimated total intracranial volume have significant impacts on the prediction of dementia.



Model 5: Ensemble Model

We have built four models in total: logistic classification, random forest, gradient boosting and extreme gradient boosting. Taking the majority vote of the result of these four models gives us the fifth model.

The miclassification error is about 0.143. The confusion matrix is shown below.

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

Final Model & Validation

We now compare the testing error and the F1 score of these five models.

Model	Test.Error	F1.Score
Logistic Model	0.198	0.780
Random Rorest	0.143	0.843
Gradient Boosting Machine	0.143	0.840
Extreme Gradient Boosting	0.143	0.847
Ensemble Model	0.143	0.843

We see that logistic regression has the highest testing error and lowest F1 Score. The other 4 models give the same testing error, but Extreme Gradient Boosting has a slightly higher F1 score.

We will still choose the ensemble model as our final model, since the ensemble method can help reduce variance and can be more robust. We use the validation dataset to evaluate our final model. The misclassification error rate turns out to be 17.4%. The model has a recall rate of 94.4% and a precision rate of 70.1%. Below, the confusion matrix suggests that the ensemble model has a high false discovery rate, which turns out to be 29.2%.

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	42	2
$\hat{Y} = 1$	14	34

VI. Conclusion

In this study, we use the longitudinal and cross-sectional MRI data of 608 elderly subjects to build different machine learning models for predicting dementia. We set the probability threshold to be 0.5 and the performance across different models are comparable. We have also used the majority vote of four models to build an ensemble model, but do not see significant improvement of results.

The Mini Mental State Exam score turns out to have high significance in all of the models we build, including logistic classification, random forest and boosting model. This suggests that a person with low mini mental state exam score will be more likely to be diagnosed with dementia.

In addition, age, nWBV (normalized whole brain volume) and eTIV (estimated total intracranial volume) also show significance. Subjects with higher age, smaller normalized whole brain volume and smaller estimated total intracranial volume have a higher chance of being diagnosed with demantia.

VII. 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/