Data Science Live 2022

Early Alzheimer's Disease Prediction

STAT 571 **Group 17**: Jia Xu, Yuqin Zhang, Zejia Cai

Wharton
University of Pennsylvania

Our Team

Jia Xu

MS Data Analytics in Social Policy ‘23

Yuqin Zhang

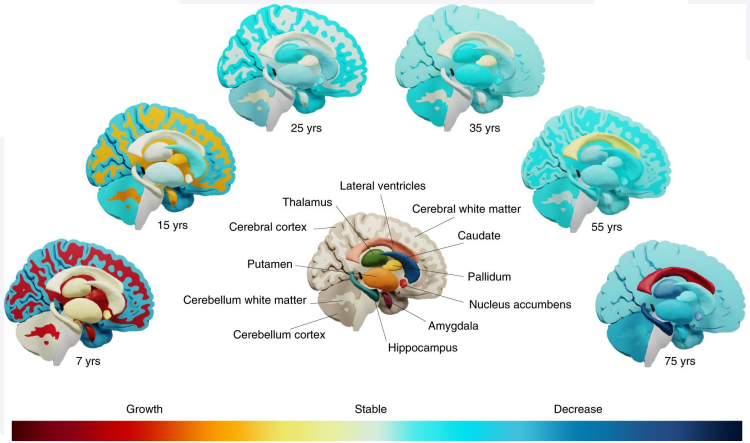
MSE Data Science ‘23

Zejia Cai

MSE Data Science ‘22

Presentation Outline

Background → Data Overview → EDA → Models → Conclusion



1 Project Background

Background Information

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.

Goal

Use Magnetic Resonance Imaging (MRI) data for both demented and nondemented adults to build classifiers that predicts whether a subject will be diagnosed to develop dementia.

2 Data Overview

Dataset Overview

- **2** Datasets: the longitudinal and cross-sectional MRI data
- **608** valid observations in total, **8** variables will be used for prediction
- **341** Non-AD observations and **267** AD observations

Demographic Variables

Gender	Gender of the subject. M = Male, F = Female.
Age	Age of the subject.
Year of Education (EDUC)	Years of education
Socioeconomic Status (SES)	A combined total measure of a person's economic and social position in relation to others

Clinical Variables

Mini Mental State Examination Score (MMSE)	A widely adopted 30-point questionnaire for measuring cognitive functions
Clinical Dementia Rating (CDR)	A global rating scale for staging patients diagnosed with dementia

Derived Anatomic Variables

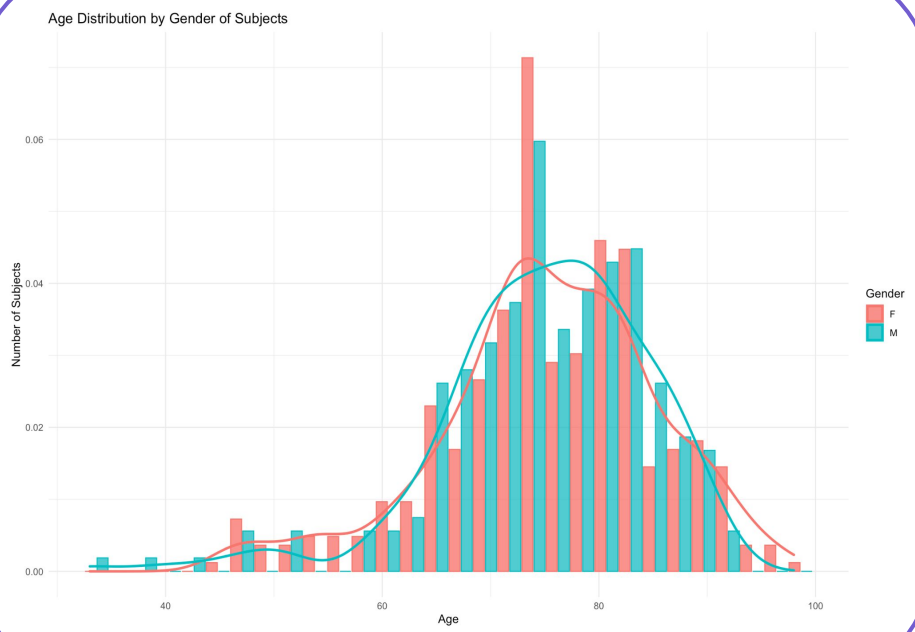
Estimated Total Intracranial Volume (eTIV)	Estimated value of the maximum pre-morbid brain volume
Normalized Whole Brain Volume (nWBV)	The percentage of brain mask occupied by voxels classified as gray and white matter
Atlas Scaling Factor (ASF)	Volume-scaling factor that standardizes the head size based on human anatomy

Target Variable

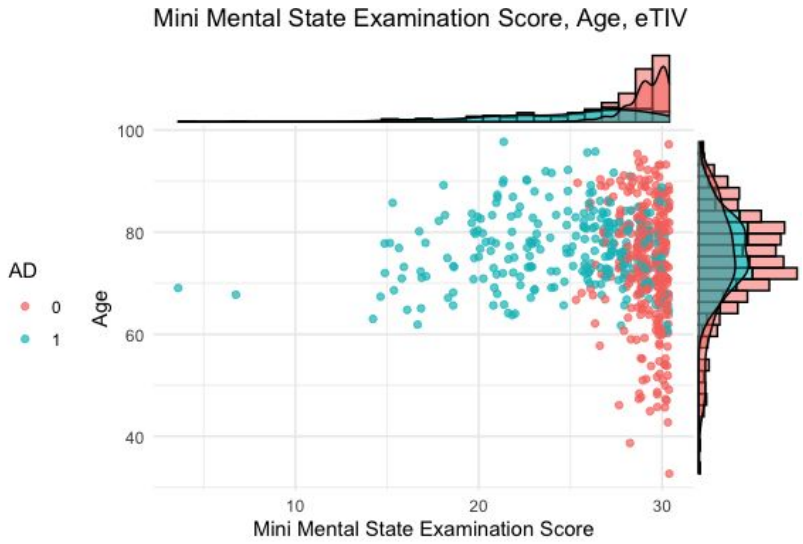
Alzheimer's Disease (AD)	Based on the value of CDR. Specifically, we let **AD = 0** for CDR = 0, and **AD = 1** for CDR > 0.

3 Exploratory Data Analysis

Age Distribution by Gender

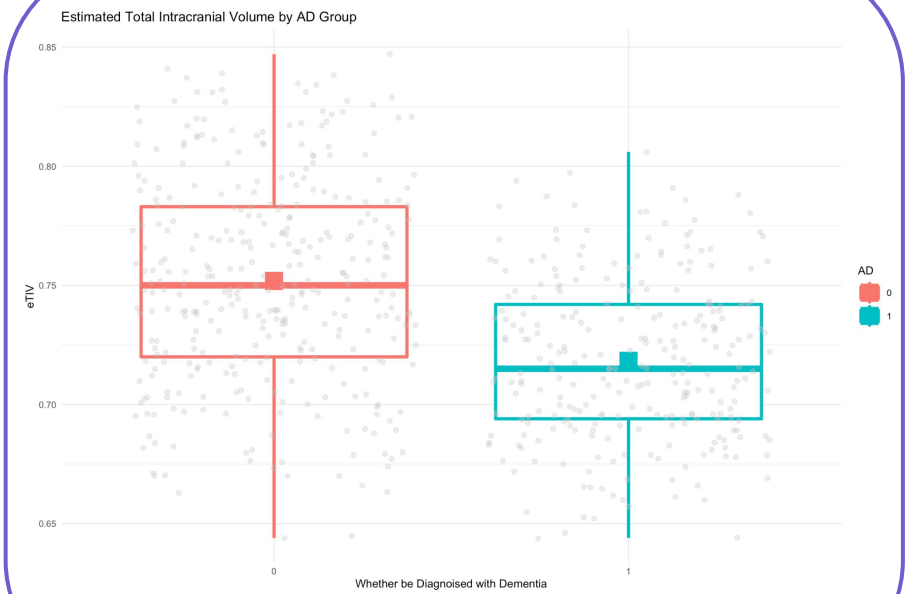


Mini Mental State Examination Score, Age and AD



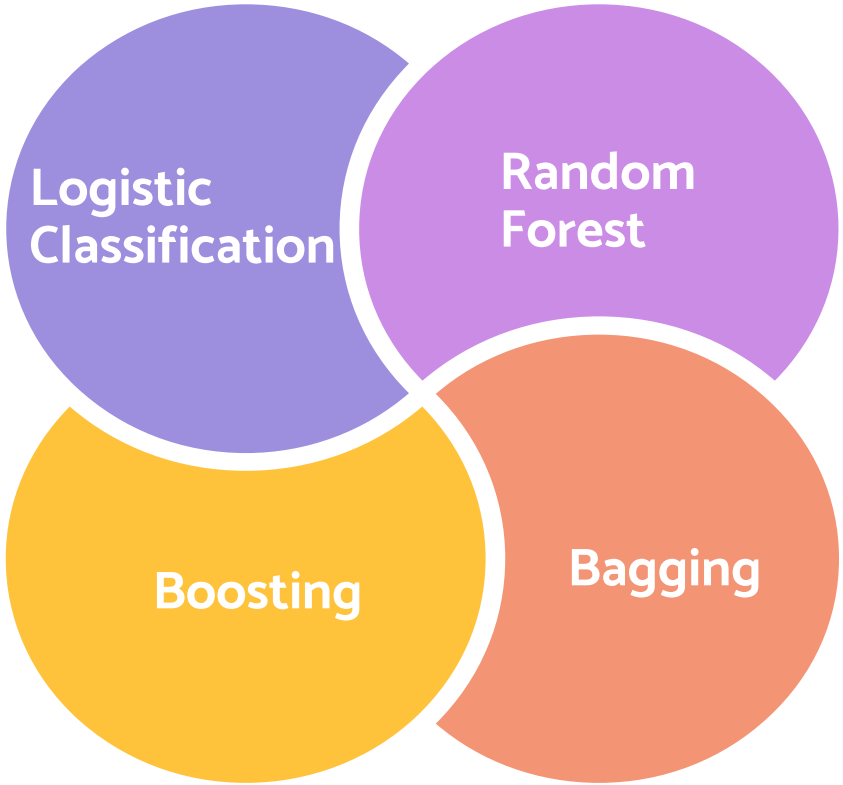
Mini Mental State Examination Score by AD Group

Estimated Total intracranial Volume by AD group



4 Models

Overview



Logistic Classification

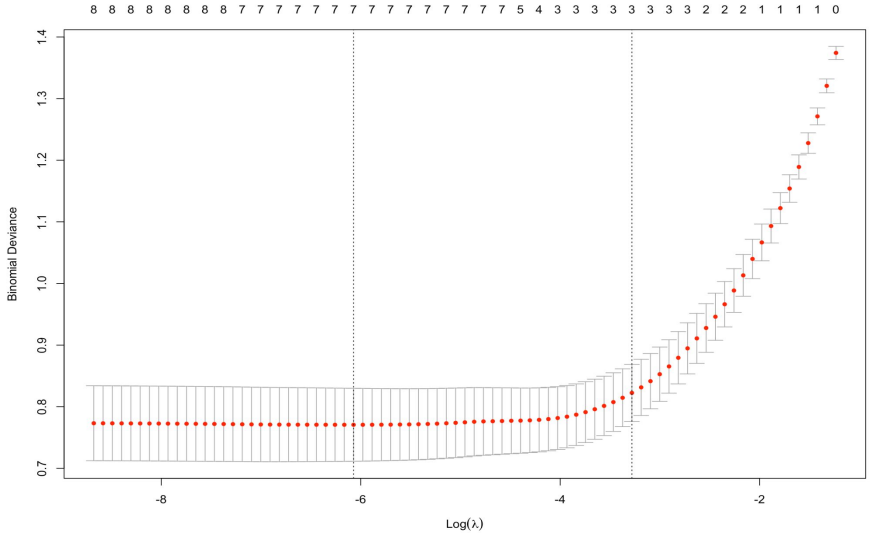
LASSO Output

Gender,
Age,
EDUC,
SES,
MMSE,
eTIV
nWBV

Backward Selection

Final Model

Gender, MMSE, nWBV



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

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

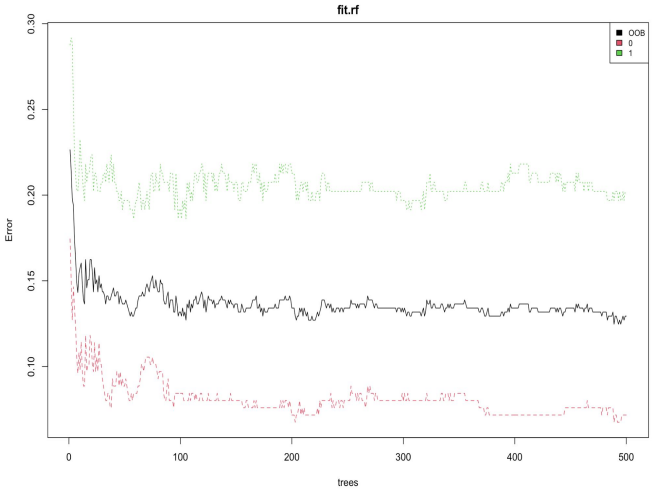
Confusion Matrix on Validation Data

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

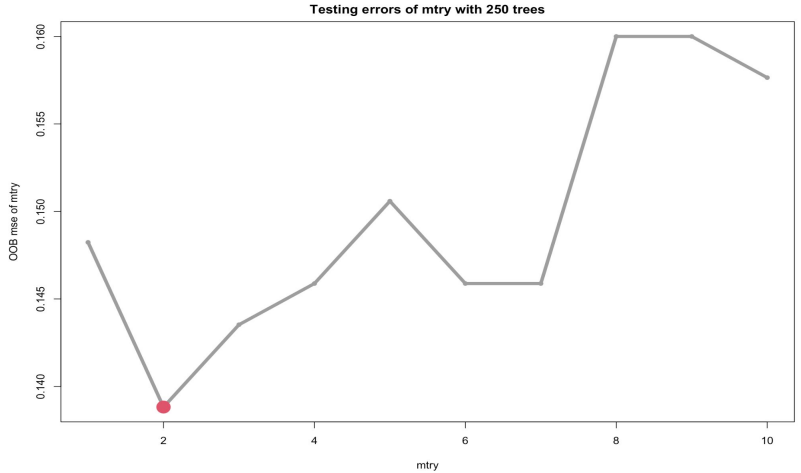
Misclassification Rate: 0.198
F1 Score: 0.780

Random Forest

Tuning **ntrees** using OOB error
Result : **250**



Tuning **mtry** using OOB error
Result: **2**



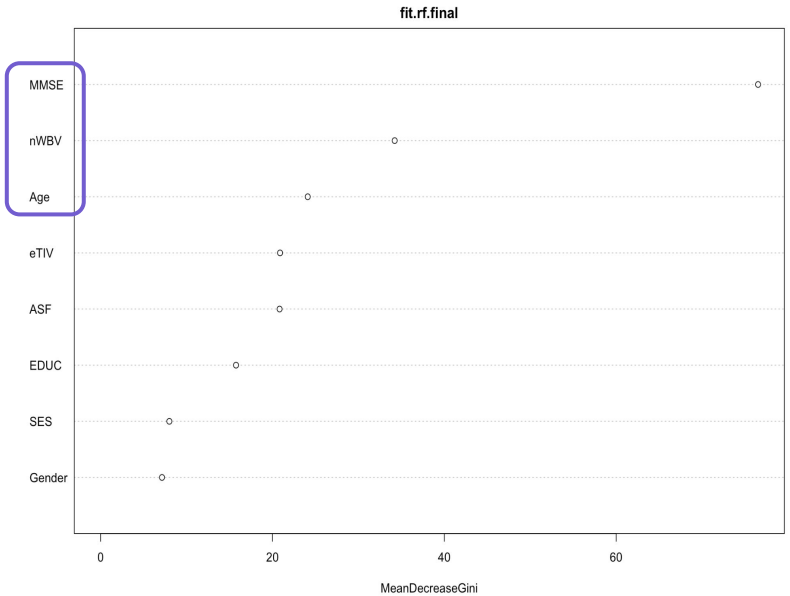
Confusion Matrix on Test Data

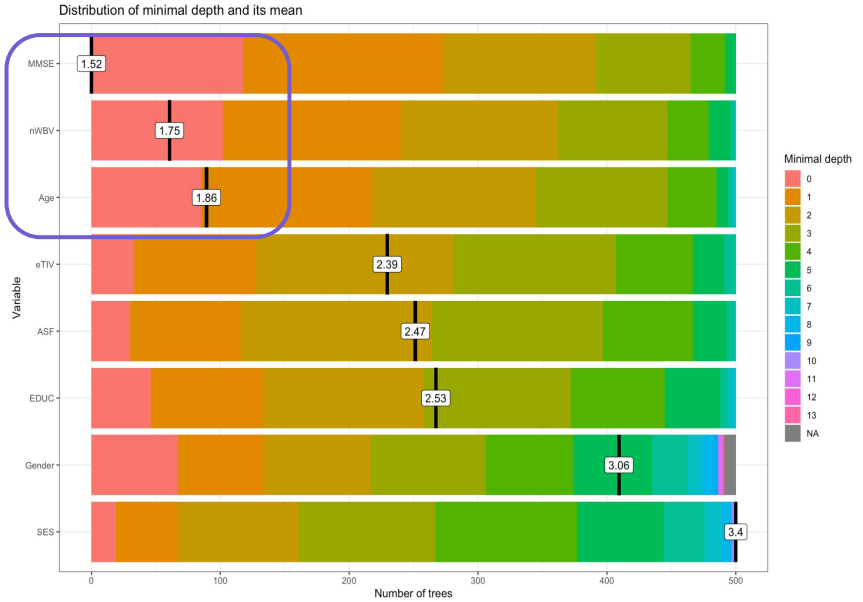
	$Y=0$	$Y=1$
$\hat{Y}=0$	43	8
$\hat{Y}=1$	5	35

Misclassification Rate: 0.143
F1 Score: 0.8433

Random Forest

Feature Importance Plots show that **MMSE, nWBV, Age** are top 3 most important features





Boosting

- Build both GBM and XGB
- Use **grid search method** to choose the best-performing set of hyperparameters

Look at **135** models for **GBM**

Parameters tuned:

- Learning rate
- Tree numbers
- Tree depth
- Minimum number of observations in the end node

Look at **240** models for **XGB**

Parameters tuned:

- Learning rate
- Tree depth
- Minimum loss reduction for a split and penalty on the number of leaves in a tree

Confusion Matrix on Test Data

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

Misclassification Rate: 0.143
F1 Score: 0.840

Confusion Matrix on Test Data

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

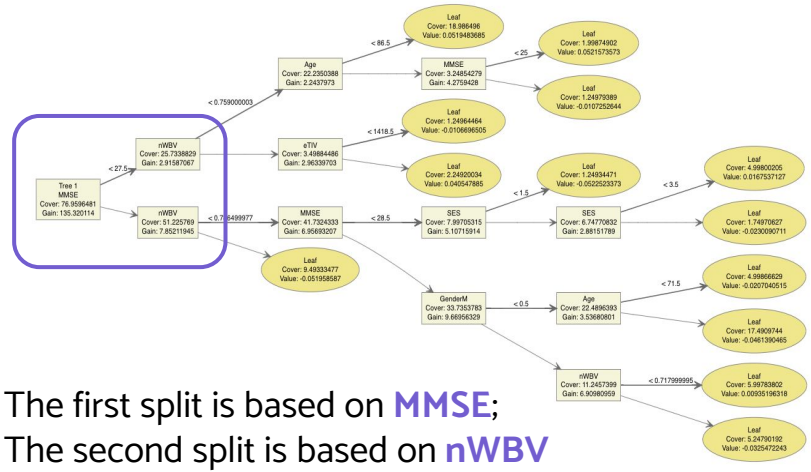
Misclassification Rate: 0.143
F1 Score: 0.847

Boosting

Feature Importance Ranking in XGB

Feature	Gain	Frequency
MMSE	**0.429**	0.134
nWBV	0.178	0.240
eTIV	0.160	0.260
Age	0.114	0.175
Educ	0.049	0.090

Display the First Tree in XGB Model



The first split is based on **MMSE**;
The second split is based on **nWBV**

Ensemble Model : Bagging

Majority Vote of Logistic Regression + Random Forest + GBM + XG Boost

(output = 1 if two or more models agree)

Confusion Matrix on Test Data

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

Misclassification Rate: 0.143
F1 Score: 0.843

Model Performance On Test Data

Model	Test Error	F1 Score
Logistic Regression	0.1978022	0.7804878
Random Forest	**0.1428571**	0.8433735
Gradient Boosting Machine	**0.1428571**	0.8395062
Extreme Gradient Boosting	**0.1428571**	**0.8470588**
Ensemble Model (all models above)	**0.1428571**	0.8433735

Final Model: Ensemble Model

Confusion Matrix on **Validation** Data

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

Misclassification Error: 17.4%

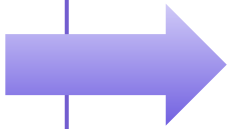
Recall: 94.4%
Precision: 70.1%

5 Conclusion

Important Factors

Top 3 Most Important Features in Each Model

- Logistic Classification:
  - Gender, MMSE, nWBV
- Random Forest:
  - MMSE, nWBV, Age
- Boosting:
  - MMSE, nWBV, eTIV



Important Features in Predicting Dementia

Demographic:
Gender, Age

Clinical:
Mini Mental State Exam Score (MMSE)

Anatomic:
Normalized Whole Brain Volume (nWBV)
Estimated Total Intracranial Volume (eTIV)

THANKS!