Group 14

TEXAS
The University of Texas at Austin

PREDICTING ALZHEIMER'S DISEASE

Aileen Li, Grace Lin, Kristen Lowe, Liyan Deng, Sidharth Saha
MSBA, The University of Texas at Austin

Agenda

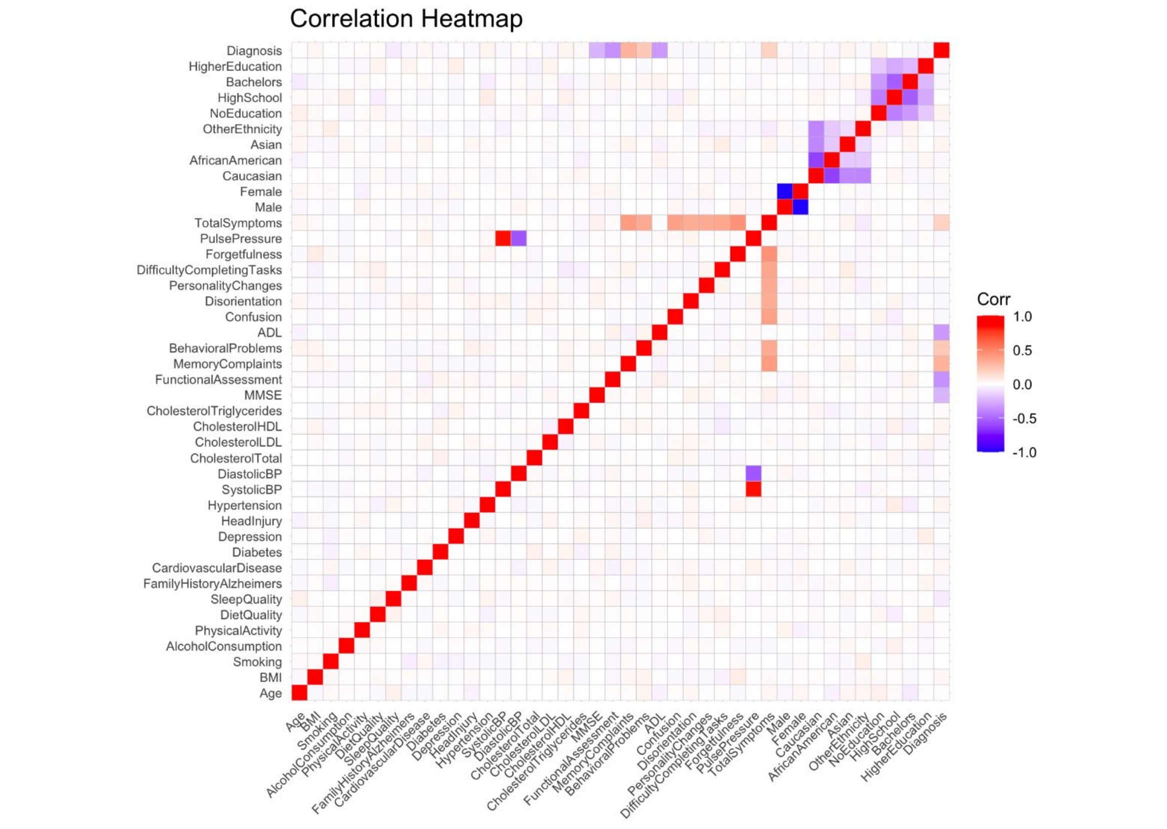
- Introduction and Exploratory Data Analysis
- KNN
- Logistic regression
- Trees (Bagging and boosting)
- Conclusion and reflection

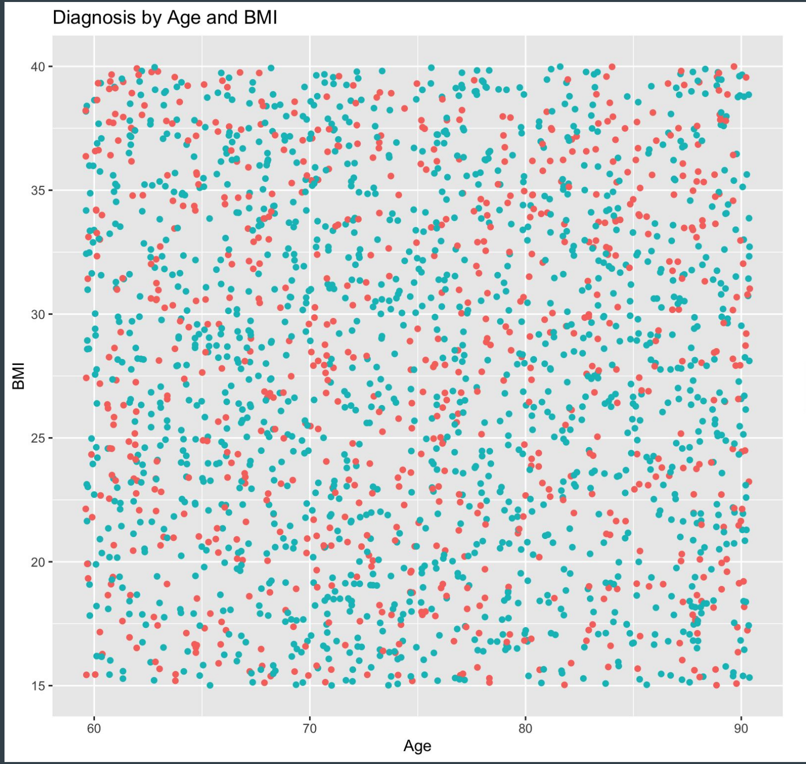
Alzheimer’s Disease Dataset

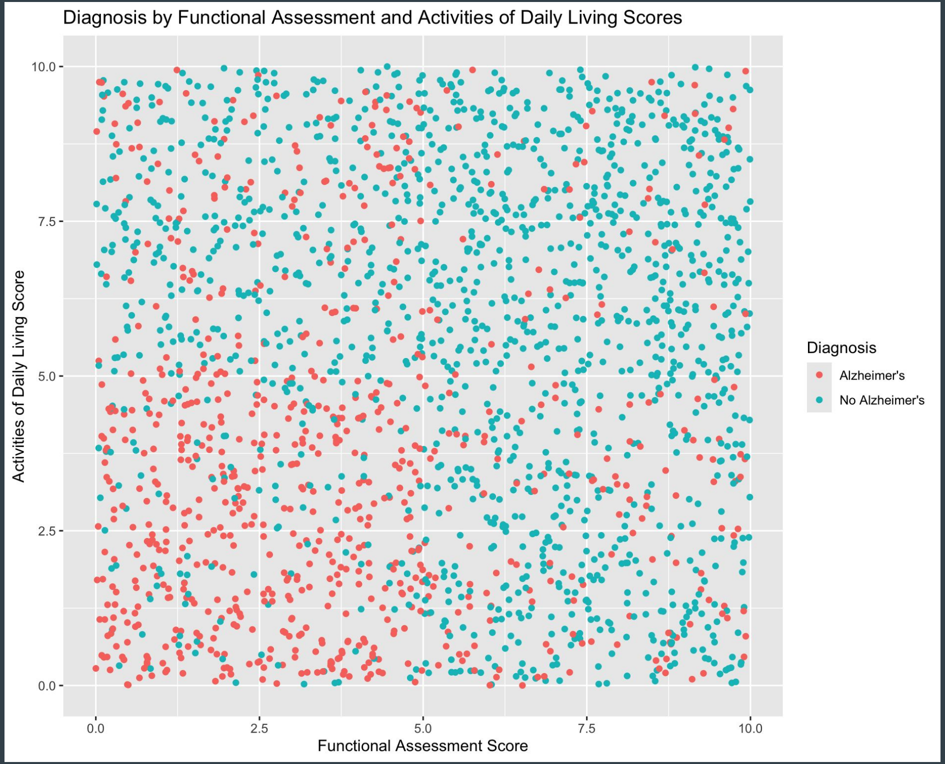
- From Kaggle
- 2,149 patient observations and 35 columns
- Target variable: Alzheimer's diagnosis (Yes/No)
- Data mildly imbalanced

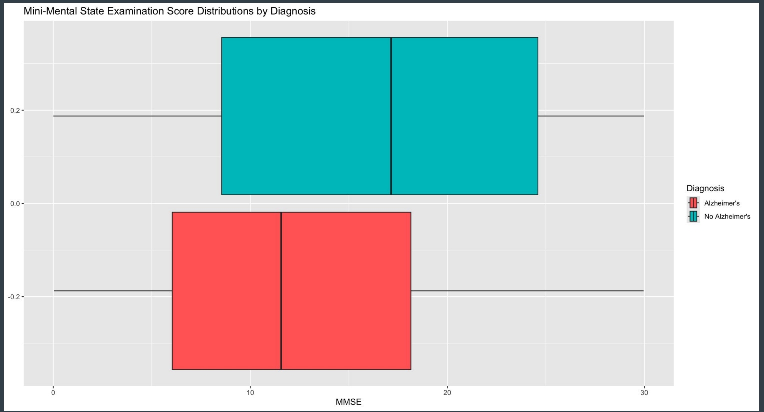
Feature Engineering

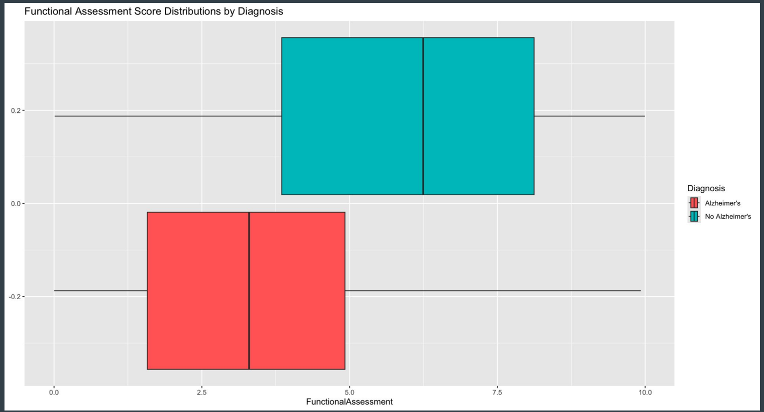
- **Made two new predictor variables**
  - Pulse Pressure
  - Total Symptoms
- **One hot encoding**
  - Gender
  - Ethnicity
  - Education Level

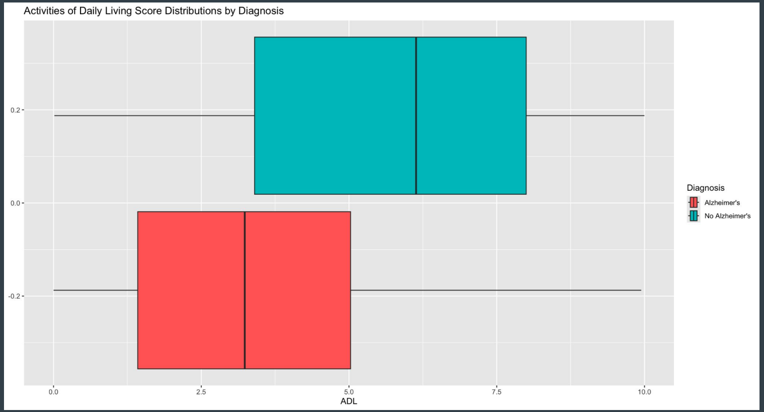

Correlation Heatmap
Diagnosis
HigherEducation
Bachelors
HighSchool
NoEducation
OtherEthnicity
Asian
AfricanAmerican
Caucasian
Female
Male
TotalSymptoms
PulsePressure
Forgetfulness
DifficultyCompletingTasks
PersonalityChanges
Disorientation
Confusion
ADL
BehavioralProblems
MemoryComplaints
FunctionalAssessment
MMSE
CholesterolTriglycerides
CholesterolHDL
CholesterolLDL
CholesterolTotal
DiastolicBP
SystolicBP
Hypertension
HeadInjury
Depression
Diabetes
CardiovascularDisease
FamilyHistoryAlzheimers
SleepQuality
DietQuality
PhysicalActivity
AlcoholConsumption
Smoking
BMI
Age
Corr
1.0
0.5
0.0
-0.5
-1.0

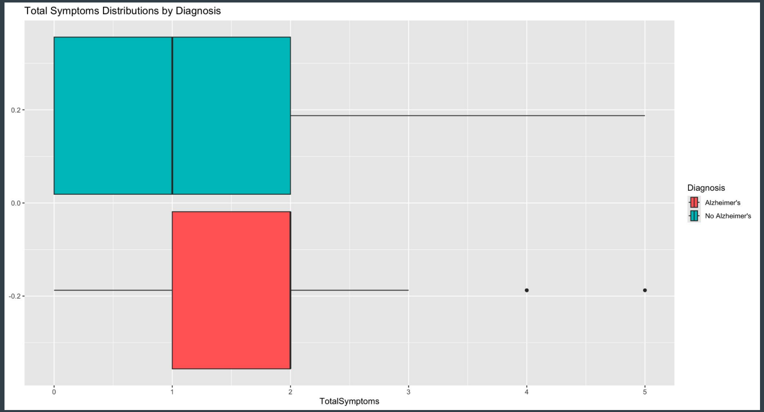

Diagnosis by Age and BMI
BMI
Age
40
35
30
25
20
15
60
70
80
90


Diagnosis by Functional Assessment and Activities of Daily Living Scores
Activities of Daily Living Score
Functional Assessment Score
10.0
7.5
5.0
2.5
0.0
Diagnosis
Alzheimer's
No Alzheimer's


Mini-Mental State Examination Score Distributions by Diagnosis
MMSE
Diagnosis
Alzheimer's
No Alzheimer's

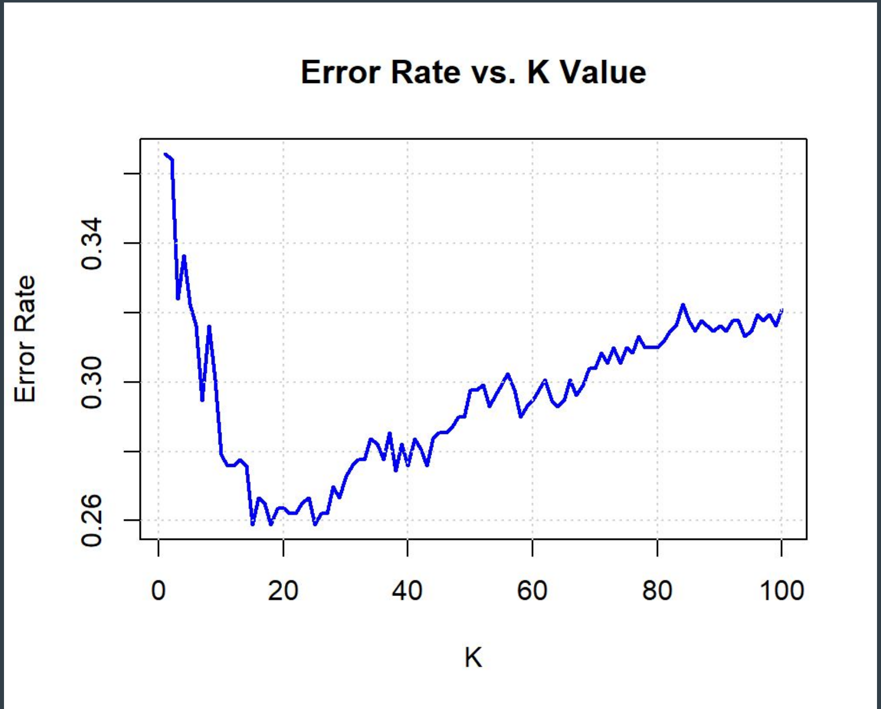

Functional Assessment Score Distributions by Diagnosis
FunctionalAssessment
Diagnosis
Alzheimer's
No Alzheimer's


Activities of Daily Living Score Distributions by Diagnosis
ADL
Diagnosis
Alzheimer's
No Alzheimer's


Total Symptoms Distributions by Diagnosis
TotalSymptoms
Diagnosis
Alzheimer's
No Alzheimer's

KNN

- Utilized all features
- Seed = 1
- k = 24



KNN Performance

- Confusion Matrix

Accuracy: 73%

		Test Data	
		YES	No
Prediction	YES	262	112
	No	6	50

Logistic Regression

- **Methods**
  - All predictors
  - All statistically significant predictors
  - Lasso
  - Ridge
  - Stepwise

Logistic Regression

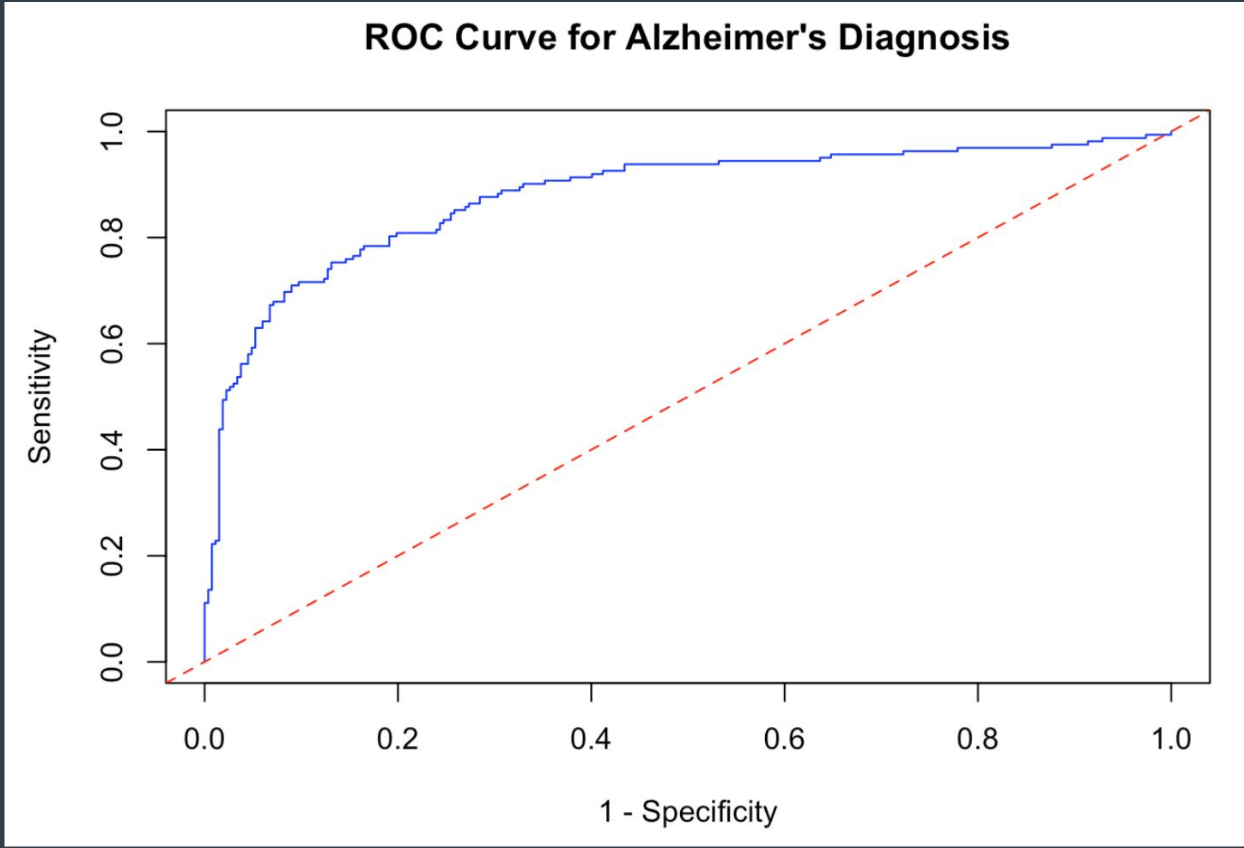
- **Use the following predictors:**
  - MMSE
  - Functional Assessment
  - Memory Complaints
  - Behavioral Problems
  - ADL
- Threshold = 0.5

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.95553  -0.27230  -0.04281   0.26716   1.14373
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.9694298  0.0267783   36.20   <2e-16 ***
## MMSE                 -0.0142408  0.0009952  -14.31   <2e-16 ***
## FunctionalAssessment -0.0536886  0.0029883  -17.97   <2e-16 ***
## MemoryComplaints      0.3343039  0.0214397   15.59   <2e-16 ***
## BehavioralProblems    0.3368824  0.0235439   14.31   <2e-16 ***
## ADL                  -0.0523070  0.0029245  -17.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Logistic Performance

Accuracy = ~83%

		Test Data	
		YES	No
Prediction	YES	110	21
	No	52	246


ROC Curve for Alzheimer's Diagnosis
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0

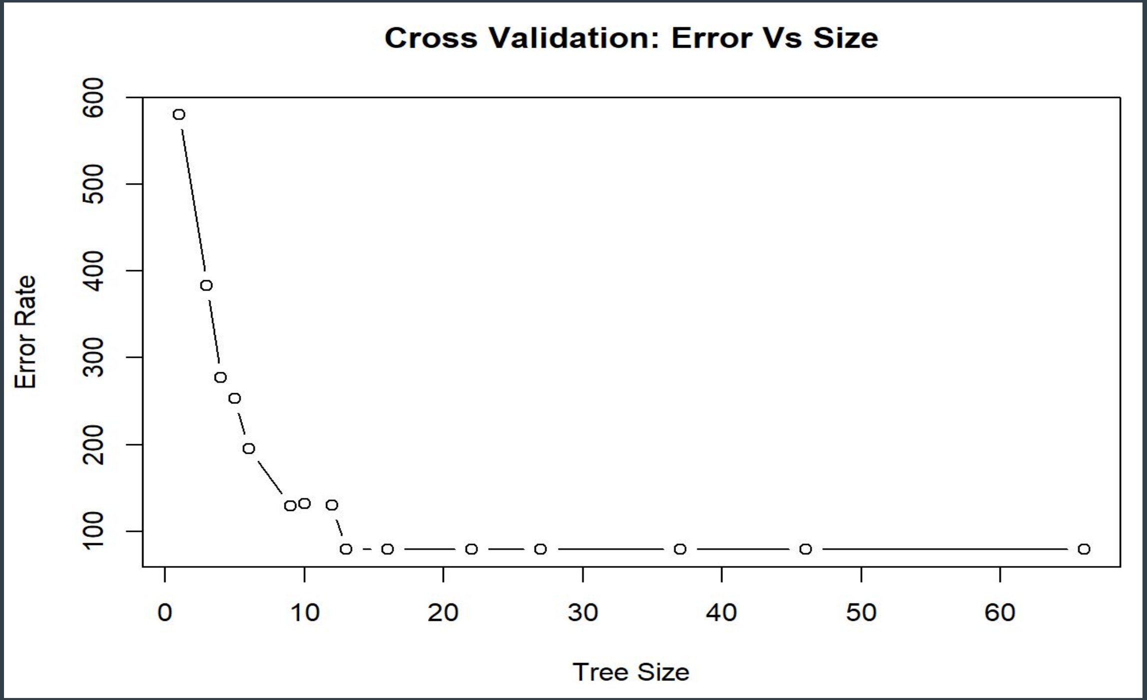
AUC = 87.89%

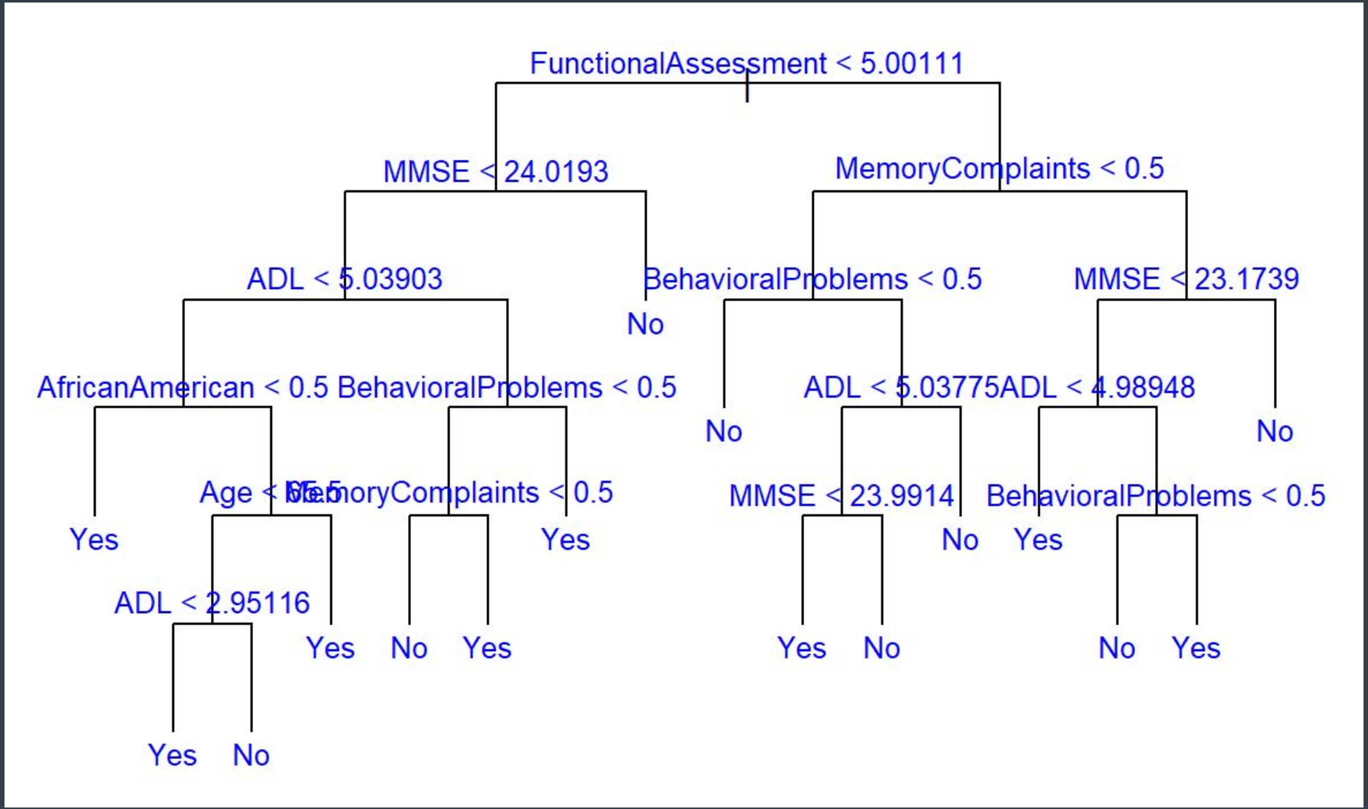
Trees

- A Classification Tree
- Random Forest
- Bagging
- Boosting

Classification Tree

- Grow & Prune method




FunctionalAssessment < 5.00111
MMSE < 24.0193
MemoryComplaints < 0.5
ADL < 5.03903
No
BehavioralProblems < 0.5
MMSE < 23.1739
AfricanAmerican < 0.5
BehavioralProblems < 0.5
No
ADL < 5.03775
ADL < 4.98948
No
Yes
Age < 85.5
MemoryComplaints < 0.5
Yes
MMSE < 23.9914
No
BehavioralProblems < 0.5
Yes
ADL < 2.95116
Yes
No
Yes
Yes
No
No
Yes
Yes
No

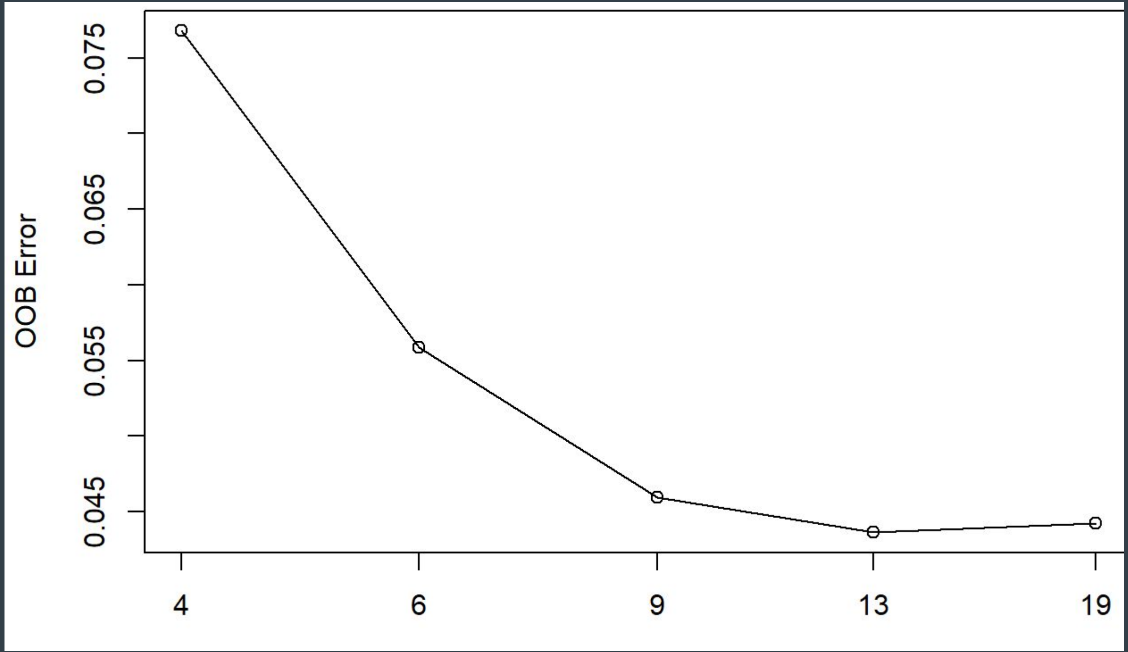
Tree Performance

Accuracy = ~94%

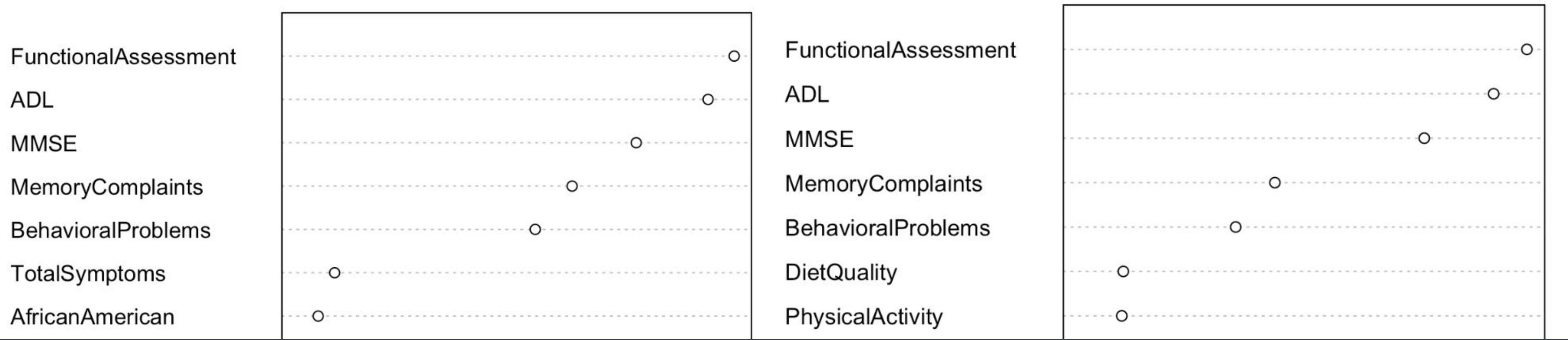
Pruned Tree		Test Data	
		YES	No
Prediction	YES	145	9
	No	17	259

Random Forest

- RFM1 : mtry = 6
- mtry tuning
- ntrees = 500



Feature Importance

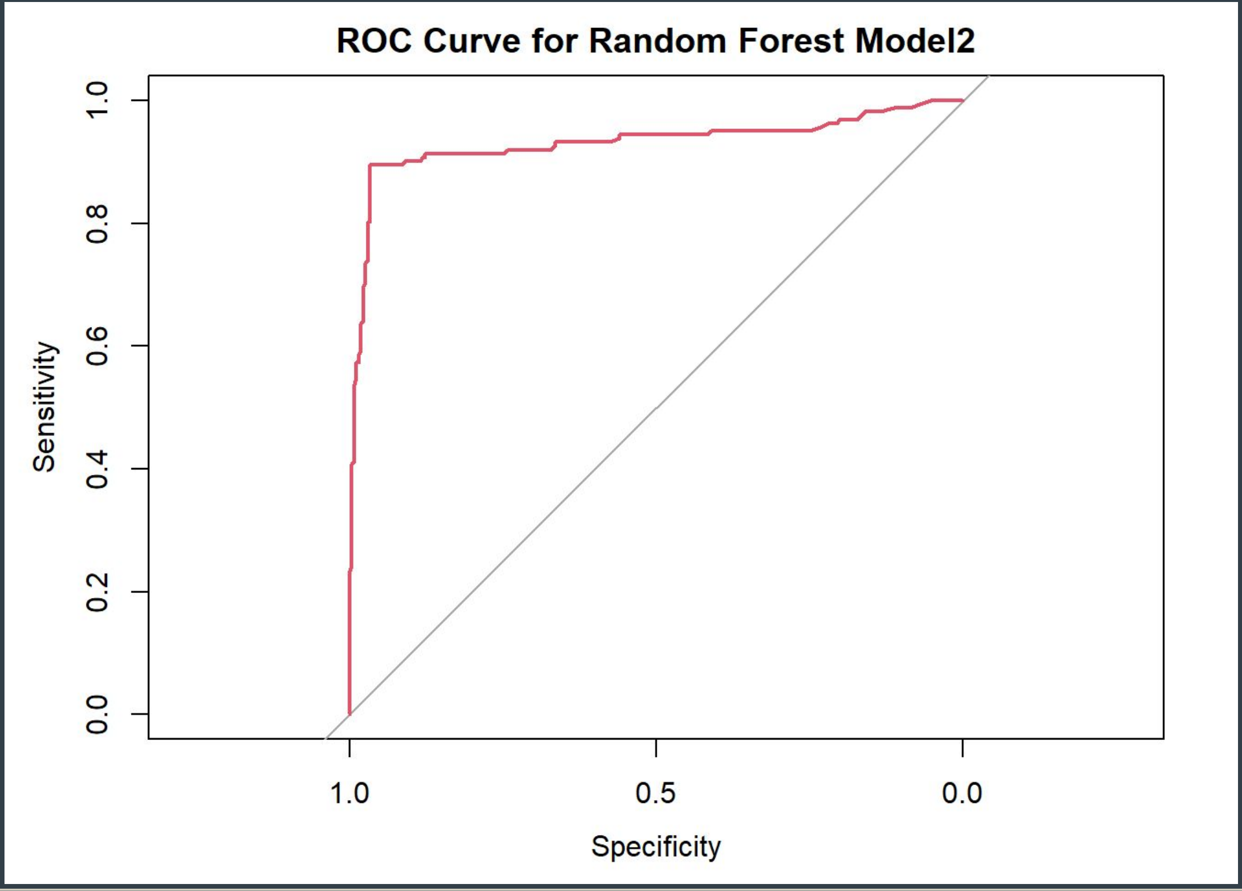


- Mean Decrease Accuracy
- Mean Decrease Gini

RF Performance

Accuracy = 95.6%*

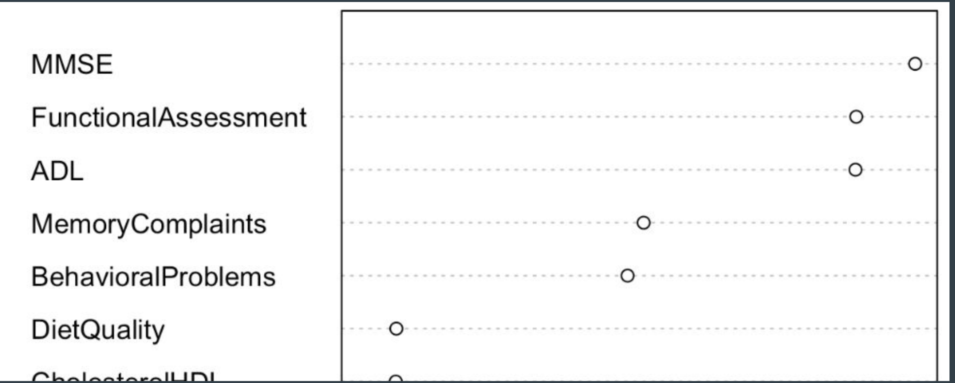
RF		Test Data	
		YES	No
Prediction	YES	549	49
	No	27	1094


ROC Curve for Random Forest Model2
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.5
0.0

Bagging

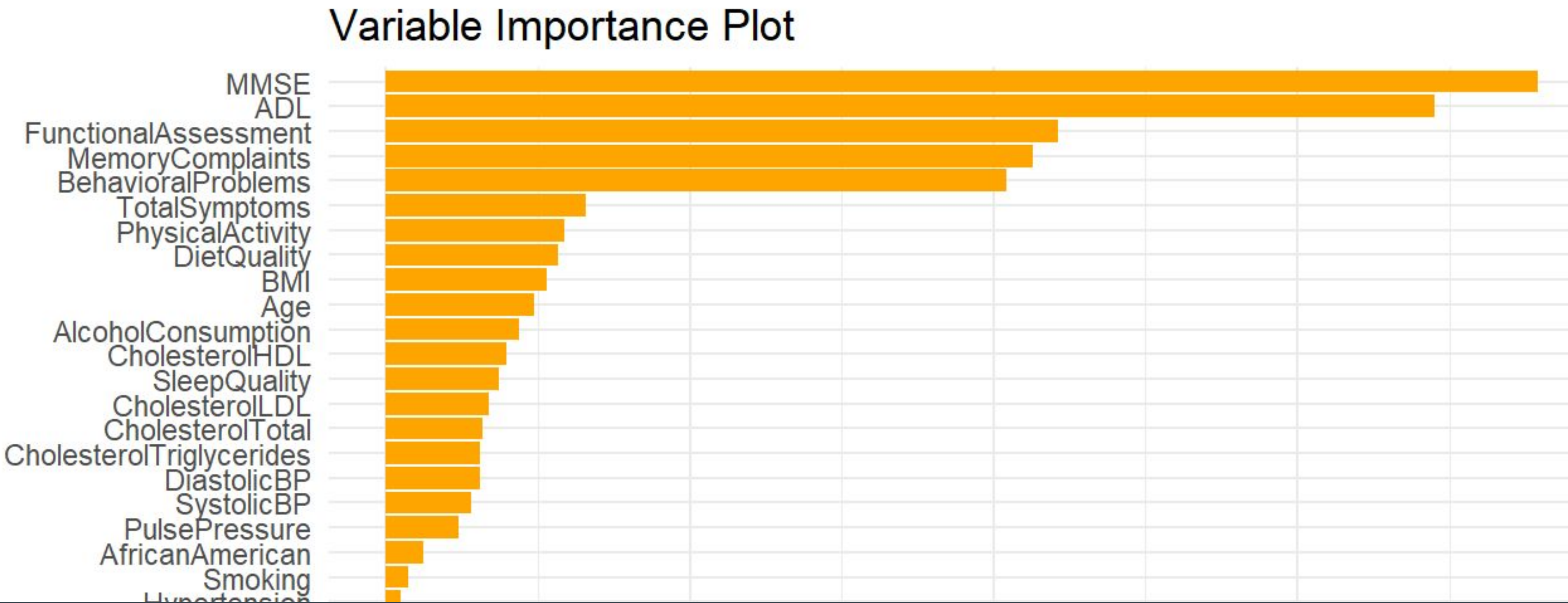
- Cross-validation
- folds = 5

Bagging Feature Importance



- Mean Decrease Accuracy
- Mean Decrease Gini

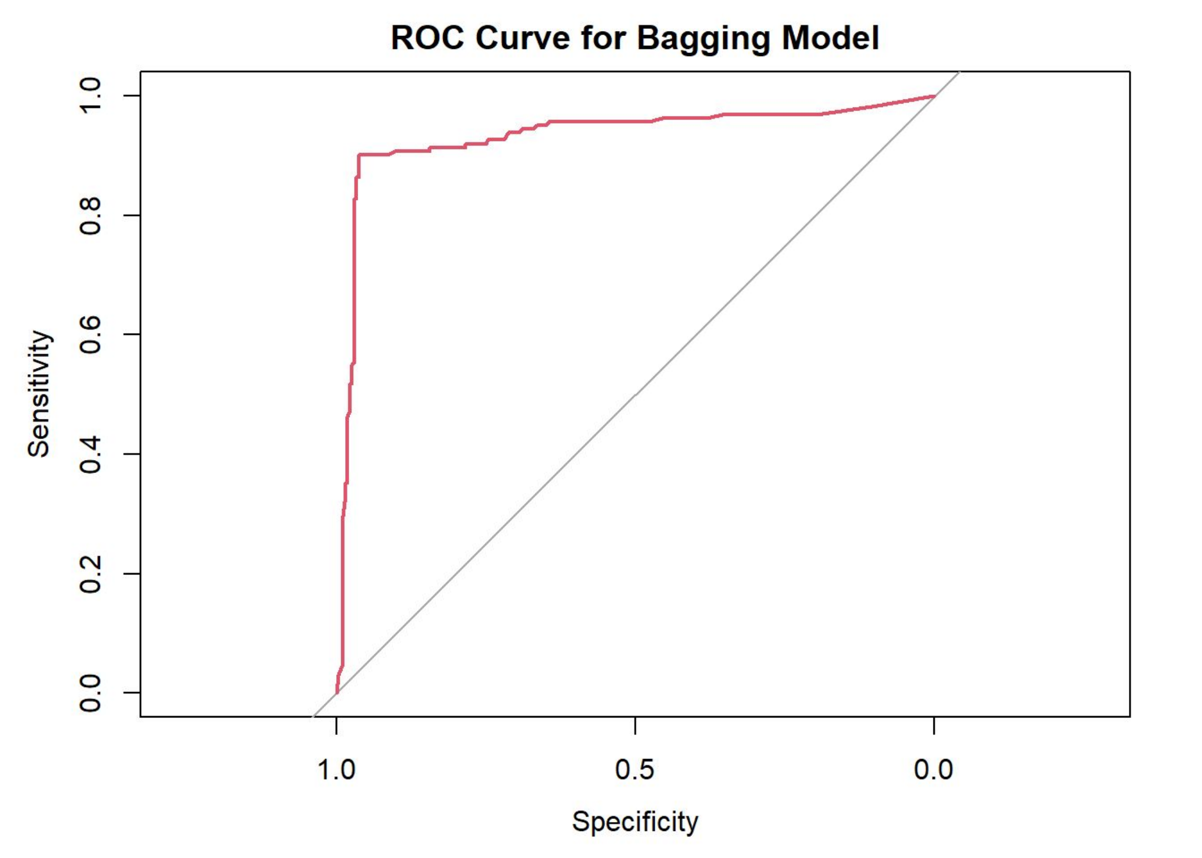
Bagging


Variable Importance Plot
MMSE
ADL
FunctionalAssessment
MemoryComplaints
BehavioralProblems
TotalSymptoms
PhysicalActivity
DietQuality
BMI
Age
AlcoholConsumption
CholesterolHDL
SleepQuality
CholesterolLDL
CholesterolTotal
CholesterolTriglycerides
DiastolicBP
SystolicBP
PulsePressure
AfricanAmerican
Smoking

Bagging Performance

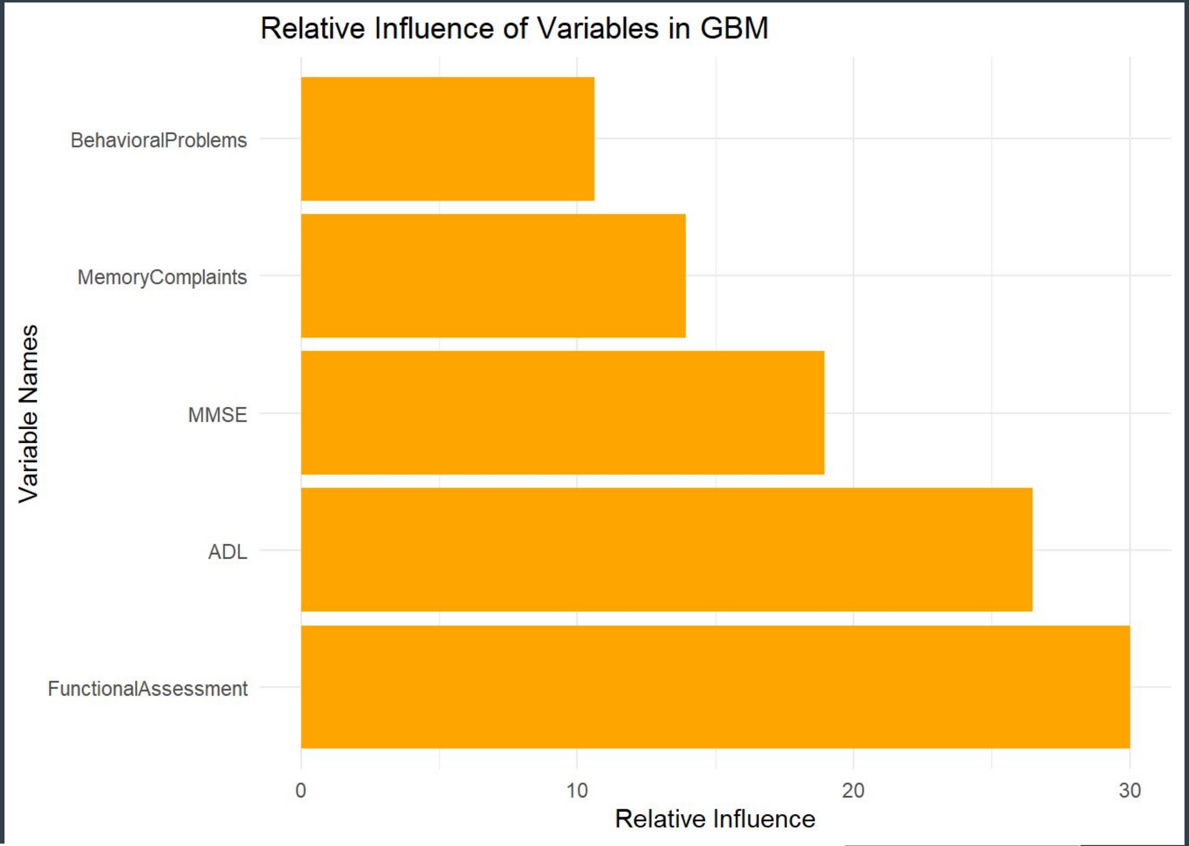
Accuracy = 92.6%

Bagging with CV		Test Data	
		YES	No
Prediction	YES	141	11
	No	21	257


ROC Curve for Bagging Model
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity

Boosting

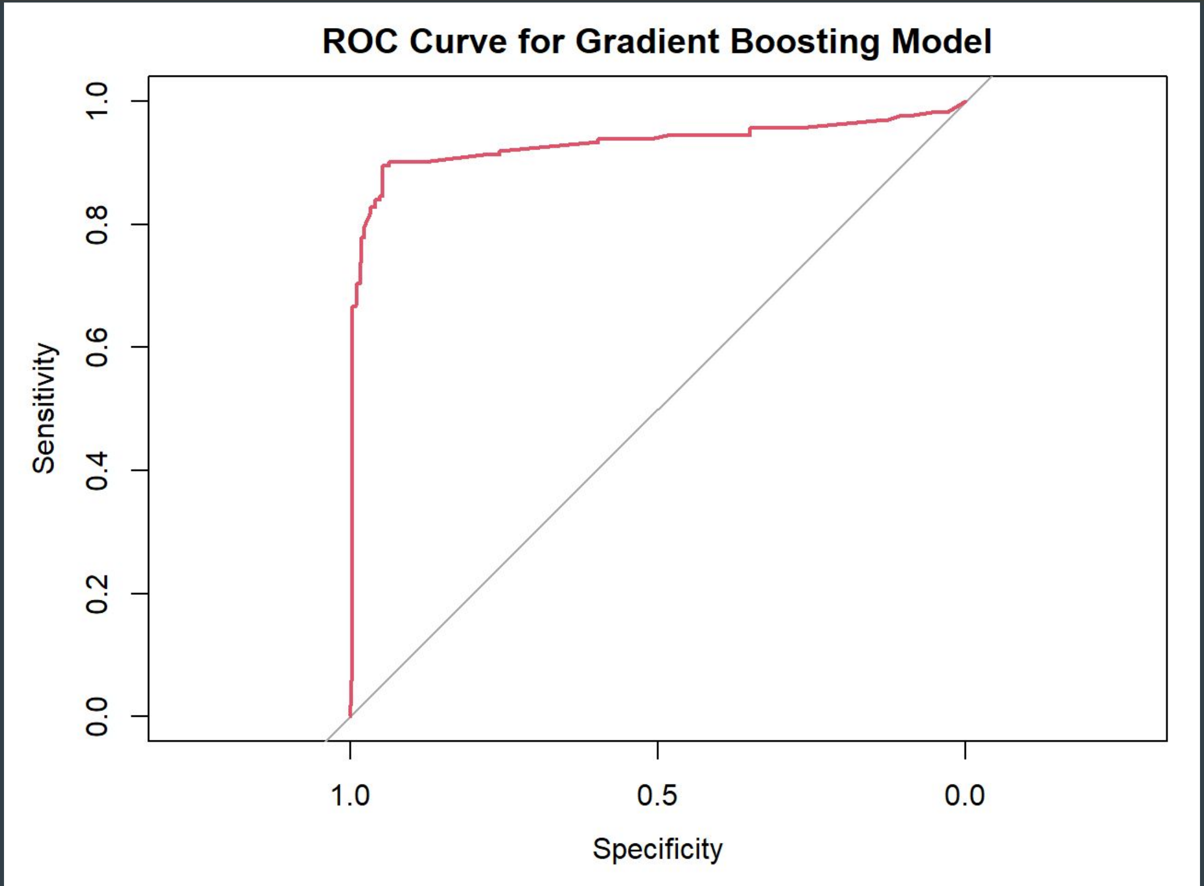
- cv =10
- n.trees = 500



Boosting

Accuracy = 91.2%

Boosting with CV		Test Data	
		YES	No
Prediction	YES	133	29
	No	9	259


ROC Curve for Gradient Boosting Model
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity

Conclusion

- Best model: **Random Forest**
- High accuracy (95.6%)
- Robustness
  - Sensitivity: 95.31%
  - Specificity: 95.71%
  - Good for imbalanced data

Reflection

- **Identifying causes/outcome relationships**
  - Feature engineering
  - Correlation analysis
- **Evaluation metrics**
  - Opportunity cost for ‘False positive’ vs ‘False negative’

Descriptive Analysis	Diagnostics Analysis
Predictive Analysis	Prescriptive Analysis

Questions?