Alzheimer's Disease/Dementia Classification Model

Group 2:
Kristine Carnavos
Edgar Lorenzo
Rahul Sinha
Mary Mulrooney

Agenda

- Background
  - Alzheimer's Disease Background
  - Data Set Summary
- Goals of Analysis
- Data Cleaning
- Data Visualization
- Data Tools/Techniques
- Results
- Conclusions

Alzheimer’s Disease Background

5.8 MILLION AMERICANS ARE LIVING WITH ALZHEIMER'S.

Every **65 SECONDS** SOMEONE IN THE UNITED STATES DEVELOPS THE DISEASE.

ALZHEIMER'S DISEASE IS THE **6TH LEADING** CAUSE OF DEATH IN THE UNITED STATES.

1 IN 3 SENIORS DIES WITH ALZHEIMER'S

ONLY 16% OF SENIORS RECEIVE REGULAR COGNITIVE ASSESSMENTS DURING ROUTINE HEALTH CHECK-UPS

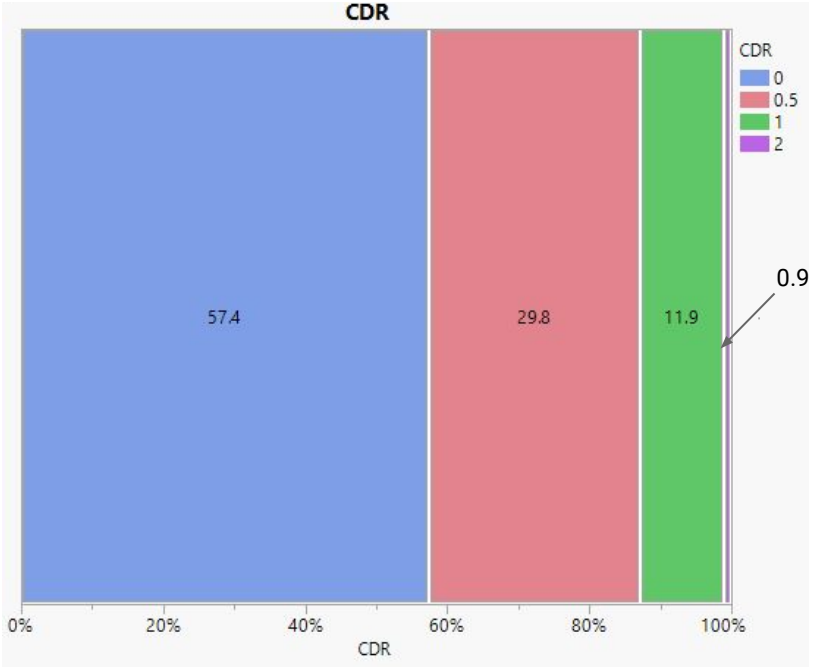
IN 2019, ALZHEIMER'S AND OTHER DEMENTIAS WILL COST THE NATION **$290 BILLION**. BY 2050, THESE COSTS COULD RISE AS HIGH AS $1.1 TRILLION.

THE LIFETIME COST OF CARE FOR AN INDIVIDUAL LIVING WITH DEMENTIA IN 2018 WAS **$350,174**.

Early diagnosis and intervention through various therapies has been shown to lessen the progression and severity of the disease→ only achieved through expensive, rigorous & time consuming testing

Data Set Summary- *MRI and Alzheimer's Disease*

- Data set obtained from Kaggle.com from the Open Access Series of Imaging Studies (OASIS)
- MRI Scans were conducted on 436 subjects ages 18 to 96, containing both male and female subjects, all right handed
- 100 of the subjects were clinically diagnosed with very mild to moderate Alzheimer's Disease
- **<u>Target variable</u>: CDR (Clinical Dementia Rating)**
  - Nominal variable
  - **Measured on a scale of 0 to 2**:
    - 0: no dementia
    - 0.5: very mild AD
    - 1: mild AD
    - 2: moderate AD



<u>Note</u>: Since values of CDR=2 encompass only 0.9% of data (2 lines), excluding these lines of data from classification model

Data Set Summary– *Predictor Variables*

- **Gender (M/F)**- nominal
- **Age**- continuous
- **Education Level**- nominal
  - Measured on scale of 1-5
- **Socioeconomic Status (SES)**- nominal
  - Measured on scale of 1-5 where 1 is highest and 5 is lowest
- **Mini Mental State Examination (MMSE)**- continuous
- **Estimated Total Intracranial Volume (eTIV)**- continuous
- **Normalized Whole Brain Volume (nWBV)**- continuous
- **Atlas Scaling Factor (ASF)**- continuous

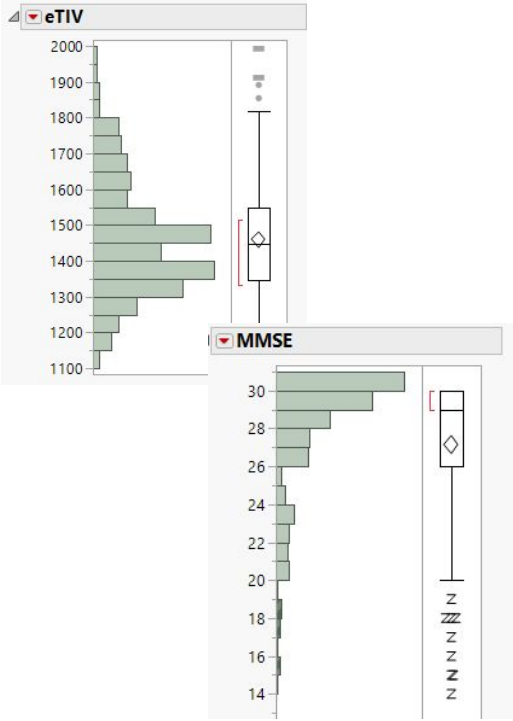
Goals of Analysis:

Determine classification model able to diagnose level of severity of Alzheimer's to allow for correct therapies & treatments to be prescribed to patients

1. **Visualize data** to understand if any obvious correlations exist for data to CDR
2. Understand **relationships of predictor variables to CDR**
3. **Formulate models** with variables of significance to try **to classify CDR**
4. **Downselect best model** to classify CDR & make recommendations for future model improvement

Data Set Cleaning- *Missing Data & Outliers*

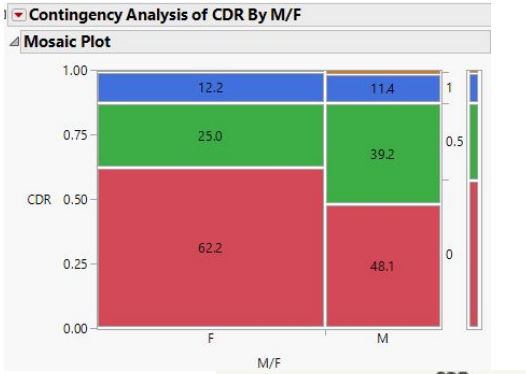
	Count	Number of columns missing	Patterns	ID	M/F	Hand	Age	Educ	SES	MMSE	CDR	eTIV	nWBV	ASF
1	216	0	000000000000	0	0	0	0	0	0	0	0	0	0	0
2	19	1	000001000000	0	0	0	0	0	1	0	0	0	0	0
3	191	4	000011110000	0	0	0	0	1	1	1	1	0	0	0

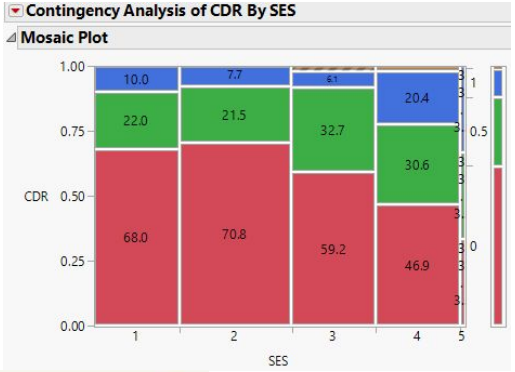


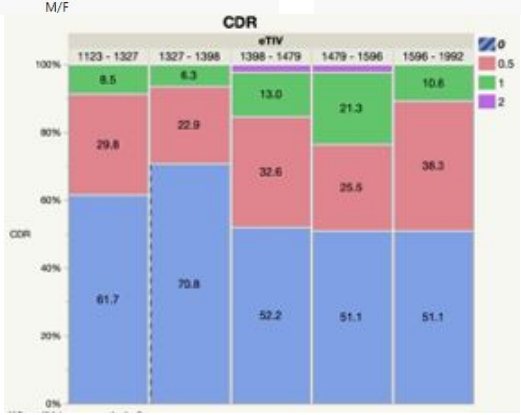
- **191 of the 436 subjects were missing both the target variable, CDR, as well as SES, Education, and MMSE**
    - To deal with the large amount of data missing from these lines, these rows of data were excluded from our analysis
- **SES** noted to be missing 19 lines of data from remaining 245 lines of data (**8%**), ignored because nominal data does not allow imputing
- Outliers for variable **eTIV** and **MMSE** were observed but encompassed **less than 5%** of the remaining 226 lines of data→ ignored

Data Visualization: *Graphical Analysis*

- Predictor variables were visualized to understand effects on CDR. The following was observed:
  - For variables **Gender, SES, and eTIV**, no difference was noted between the measured CDR for each of the groups (p-values of 0.1361, 0.2522, and 0.2540)

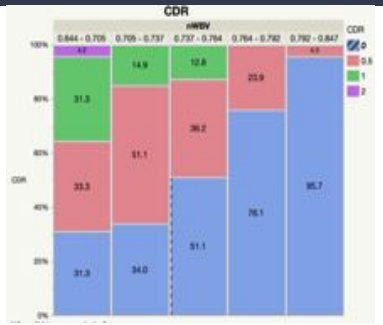


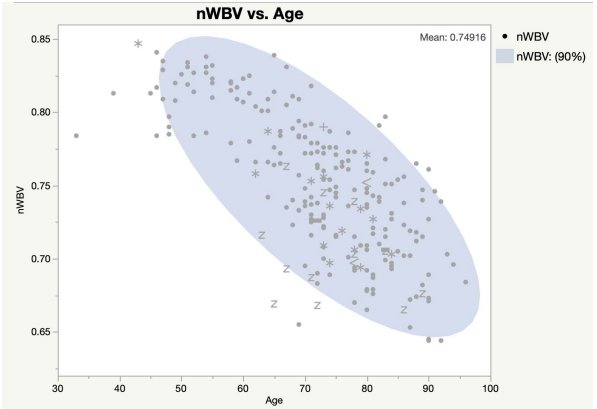


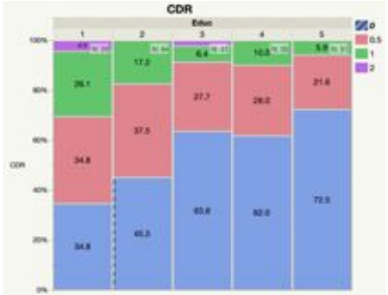


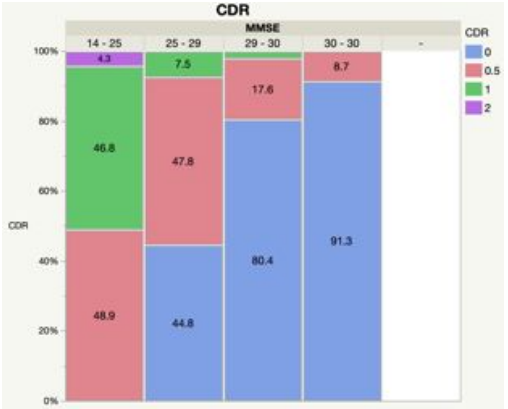
Data Visualization: *Graphical Analysis*

- **Age, nWBV, and MMSE** had significant impact on CDR (p-values < 0.0001)
- Correlation between nWBV and Age
- Greater educational attainment was associated with a lower likelihood of dementia due to Alzheimer's









Data Tools & Techniques: *Correlations & PCA*

Correlations

	CDR	Age	Educ	SES	MMSE	eTIV	nWBV	ASF
CDR	1.0000	0.3000	-0.2547	0.2144	-0.7501	0.1052	-0.5010	-0.1143
Age	0.3000	1.0000	-0.2071	0.1555	-0.2521	0.0401	-0.7203	-0.0345
Educ	-0.2547	-0.2071	1.0000	-0.7366	0.2973	0.1473	0.1949	-0.1266
SES	0.2144	0.1555	-0.7366	1.0000	-0.2599	-0.1712	-0.1345	0.1562
MMSE	-0.7501	-0.2521	0.2973	-0.2599	1.0000	-0.0057	0.4696	0.0141
eTIV	0.1052	0.0401	0.1473	-0.1712	-0.0057	1.0000	-0.2189	-0.9897
nWBV	-0.5010	-0.7203	0.1949	-0.1345	0.4696	-0.2189	1.0000	0.2192
ASF	-0.1143	-0.0345	-0.1266	0.1562	0.0141	-0.9897	0.2192	1.0000

Principal Components / Factor Analysis

Principal Components: on Correlations

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	1.9897	99.487		99.487
2	0.0103	0.513		100.000

Correlations:

- **ASF & eTIV**: Highly negatively correlated (98.97%), not strongly correlated to any other variables→ PCA Analysis Conducted for Variables
- Other strong negative correlations noted between **education and SES** and **nWBV and Age**, but due to correlations of at least 1 of each of the variables in the pair on target, PCA was not performed.

Formula for Princ1 Variable:

0.0044092741 • eTIV

+ -5.463358731 • ASF

+ 0.210907798

Classification Model Creation Summary

- **Validation**
  - 60/20/20 Split
  - Stratified Random on CDR
- **Models Run**
  - Nominal Logistic
  - Decision Tree (Partition)
  - Bootstrap Forest
  - Neural Network (One Layer)

Model Compare- *Classification Model*

Model Comparison Validation=Validation

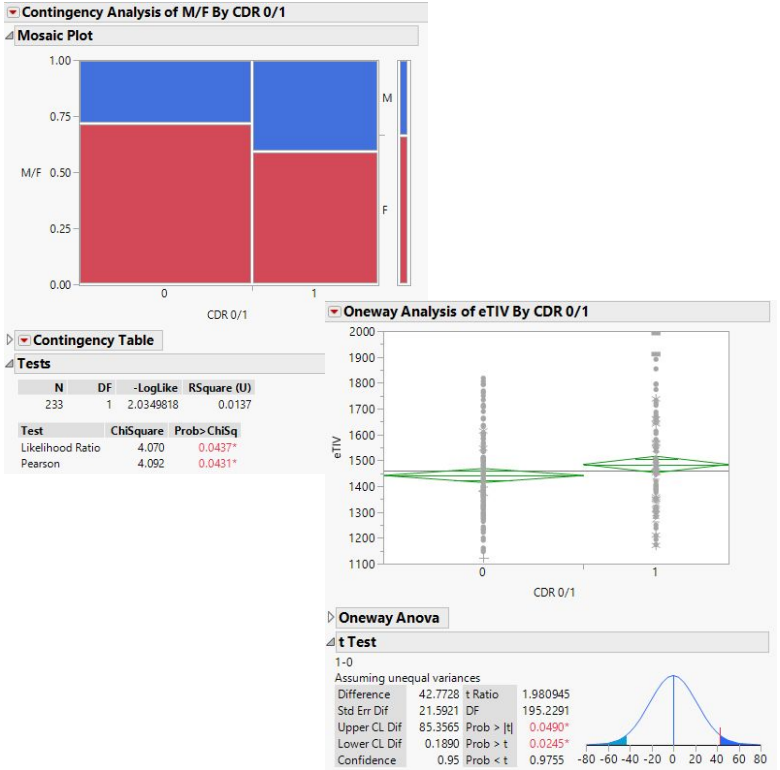
Measures of Fit for CDR

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Fit Nominal Logistic		0.4305	0.6552	0.5364	0.4086	0.3229	0.2553	47
Partition		0.3768	0.5994	0.587	0.4360	0.3623	0.2553	47
Bootstrap Forest		0.2134	0.3904	0.7409	0.5147	0.4729	0.3191	47
Neural		0.4262	0.6432	0.52	0.4145	0.3399	0.2093	43

- Of the 4 models run, the model that best predicts CDR for the validation data set is the **Neural Network**
- Neural Network has **lowest misclassification rate: 20.93%**
  - Not very good model, confusion matrices show some risk of false positives
  - Data exclusion of 201 lines missing data may have resulted in reduced ability to classify
  - Additional issue may be data set itself→ over 57% of data set was 0 (No Dementia)

Alternative Approach- *Prediction Model*

- Recoded CDR as 0/1 nominal data and run as a **prediction model**
  - **0**: Will not get Alzheimer's Disease (57.4%)
  - **1**: Will get Alzheimer's Disease (42.6%)
- <u>Benefits of Recoding Data to Prediction Model:</u>
  - Allows better distribution of target variable in data set
  - Allows additional models to be considered (ie Boosted Tree)
- <u>Noted differences between classification & prediction model</u>:
  - **Gender** and **eTIV** are noted to have statistically different means when grouped by 0/1 CDR variable (were not different in classification model)



Alternative Approach- *Prediction Model*

Model Comparison Validation- Prediction=Validation

Predictors

Measures of Fit for CDR 0/1

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Fit Nominal Logistic		0.7447	0.8564	0.1731	0.2109	0.1410	0.0435	46
Partition		0.7335	0.8489	0.1807	0.2128	0.1463	0.0435	46
Boosted Tree		0.6703	0.8043	0.2235	0.2495	0.1626	0.0652	46
Bootstrap Forest		0.3979	0.5624	0.41	0.3406	0.3316	0.0444	45
Neural		0.7396	0.8534	0.1774	0.1983	0.1502	0.0444	45

- For the 5 models run, the best predictor model was found to be **Fit Nominal Logistic**
  - **R-squared value of 74.47%**
  - Neural Network and Partition were also fairly close with RSquared Values of 73.96% and 73.35%
  - Misclassification rate of all models are fairly low as well (4.35-4.44%)
  - Still room for improvement, but points towards data being better suited for predictor model in current state for if Alzheimer's will occur rather than a classification model.

Recommendations & Conclusions

1. **Based on analysis, data set is better suited for a prediction model than a classification model**
    a. R-Squared of prediction model: 75.17%
    b. Misclassification Rate of classification model: 20.93%
2. **To improve classification abilities of the model:**
    a. CDR sample should be leveled to include more patients from 0.5 to 2 categories (currently, 57% of data set represents CDR=0)
        i. This will allow CDR=2 to be included in model (more serious form of alzheimer's).
    b. Excluded lines of data due to large amounts of missing prediction variables should be looked into to determine if those variables can be collected
    c. If large amounts of predictor variables are still missing, impute data set and rerun model to see if classification rate improves

Thank You!

Questions?

Appendix

Data Tools & Techniques - *Nominal Logistic (Classification)*

- Misclassification Rate: 24.29%
- Significant predictor variables (p<0.05):
  - MMSE
  - nWBV
- Same for if run for forward/backward analysis

Nominal Logistic Fit for CDR

Effect Summary

Source	LogWorth	PValue
MMSE	16.115	0.00000
nWBV	2.541	0.00288

Remove Add Edit Undo FDR

Converged in Gradient, 6 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	54.89309	4	109.7862	<.0001*
Full	75.83972			
Reduced	130.73281			

RSquare (U)	0.4199
AICc	164.311
BIC	181.329
Observations (or Sum Wgts)	140

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4199	0.4305	0.2404	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6428	0.6552	0.4241	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5417	0.5364	0.6958	$\sum$ -Log(ρ[j])/n
RMSE	0.4214	0.4086	0.4966	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.3290	0.3229	0.3880	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.2429	0.2553	0.3913	$\sum$ (ρ[j]≠ρMax)/n
N	140	47	46	n

Confusion Matrix

Training

Actual	Predicted Count		
CDR	0	0.5	1
0	73	8	0
0.5	15	23	4
1	1	6	10

Validation

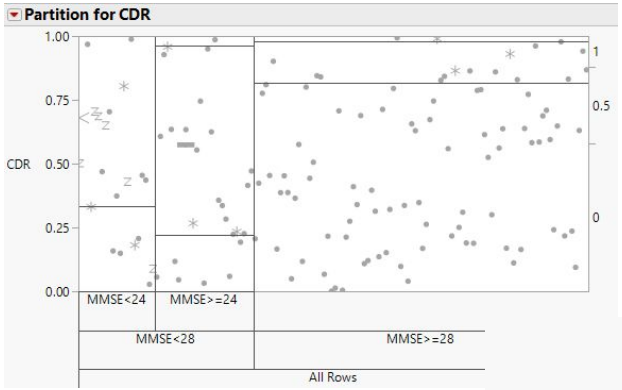
Actual	Predicted Count		
CDR	0	0.5	1
0	23	4	0
0.5	2	9	3
1	0	3	3

Test

Actual	Predicted Count		
CDR	0	0.5	1
0	23	4	0
0.5	7	3	4
1	0	3	2

Data Tools & Techniques – *Decision Tree (Classification)*

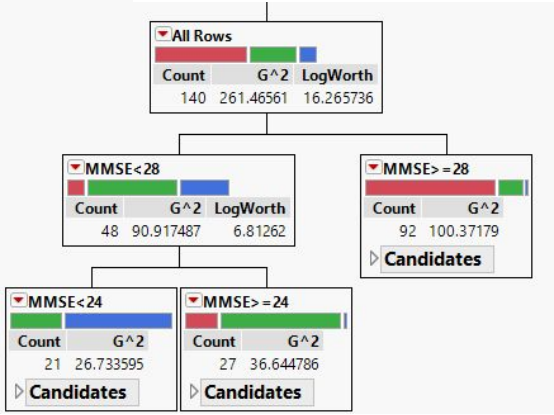
- Number of Splits: 2
- Partition Variables:
  - MMSE (<24 or >=24)
  - MMSE (<28 or >=28)
- Misclassification rate: 22.14%



	RSquare	N	Number of Splits
Training	0.370	140	2
Validation	0.377	47	
Test	0.131	46	

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3696	0.3768	0.1308	1-Loglike(mod
Generalized RSquare	0.5897	0.5994	0.2537	(1-(L(0)/L(mod
Mean -Log p	0.5887	0.5870	0.7962	∑ -Log(ρ[j])/n
RMSE	0.4298	0.4360	0.5204	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3518	0.3623	0.4328	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2214	0.2553	0.3913	∑ (ρ[j]≠pMax)/
N	140	47	46	n



Confusion Matrix

Training

Actual CDR	Predicted Count 0	0.5	1
0	75	6	0
0.5	15	20	7
1	2	1	14

Validation

Actual CDR	Predicted Count 0	0.5	1
0	24	3	0
0.5	3	6	5
1	0	1	5

Test

Actual CDR	Predicted Count 0	0.5	1
0	24	3	0
0.5	7	1	6
1	0	2	3

Data Tools & Techniques – *Bootstrap Forrest (Classification)*

- Misclassification Rate: 30%
- Number of Trees: 4

Bootstrap Forest for CDR

Specifications

Target Column:	CDR	Training Rows:	140
Validation Column:	Validation	Validation Rows:	47
		Test Rows:	46
Number of Trees in the Forest:	4	Number of Terms:	7
Number of Terms Sampled per Split:	1	Bootstrap Samples:	140
		Minimum Splits per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3038	0.2134	0.1948	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5121	0.3904	0.3573	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6501	0.7409	0.7376	$\sum$ -Log(ρ[j])/n
RMSE	0.4805	0.5147	0.5188	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.4328	0.4729	0.4774	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.3000	0.3191	0.3478	$\sum$ (ρ[j]≠ρMax)/n
N	140	47	46	n

Confusion Matrix

Training

Actual	Predicted Count		
CDR	0	0.5	1
0	78	3	0
0.5	26	15	1
1	4	8	5

Validation

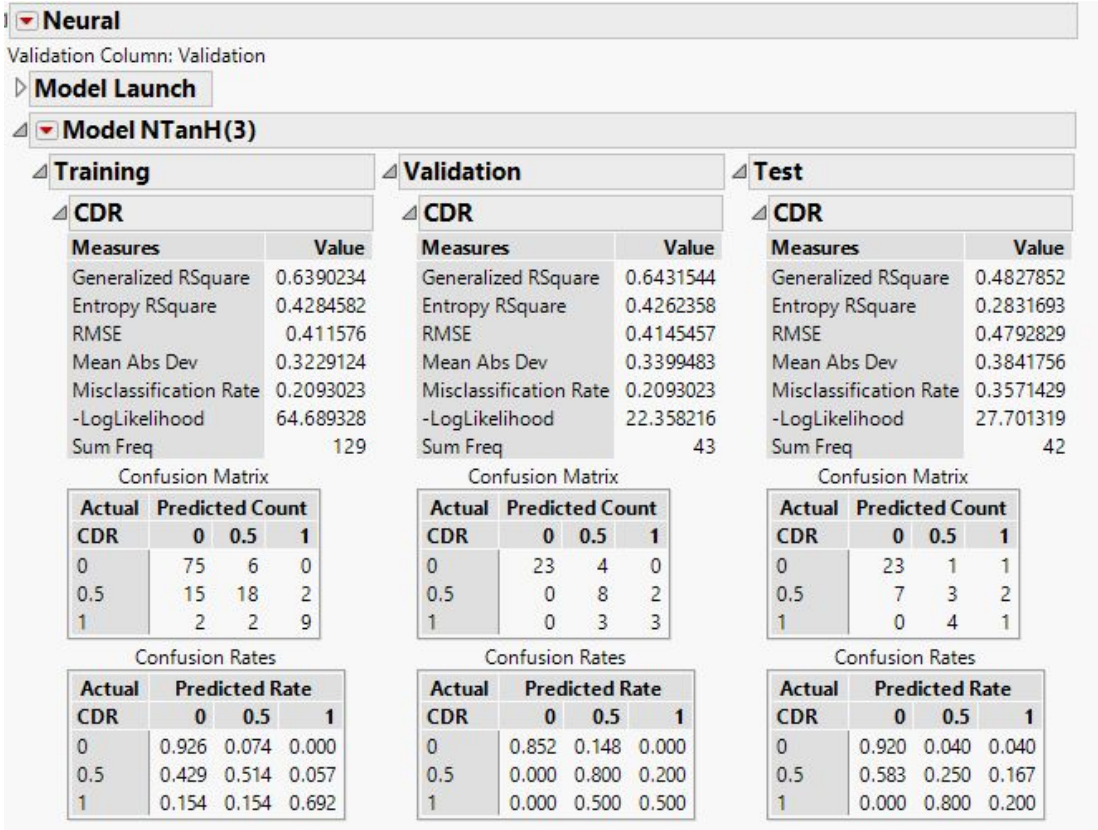
Actual	Predicted Count		
CDR	0	0.5	1
0	24	3	0
0.5	7	7	0
1	1	4	1

Test

Actual	Predicted Count		
CDR	0	0.5	1
0	26	1	0
0.5	8	4	2
1	1	4	0

Data Tools & Techniques – *Neural Network (Classification)*

- Misclassification Rate: 20.93%
- One layer used due to limited complexity of model



Neural

Validation Column: Validation

Model Launch

Model NTanH(3)

Training

CDR

Measures	Value
Generalized RSquare	0.6390234
Entropy RSquare	0.4284582
RMSE	0.411576
Mean Abs Dev	0.3229124
Misclassification Rate	0.2093023
-LogLikelihood	64.689328
Sum Freq	129

Confusion Matrix

Actual	Predicted Count		
CDR	0	0.5	1
0	75	6	0
0.5	15	18	2
1	2	2	9

Confusion Rates

Actual	Predicted Rate		
CDR	0	0.5	1
0	0.926	0.074	0.000
0.5	0.429	0.514	0.057
1	0.154	0.154	0.692

Validation

CDR

Measures	Value
Generalized RSquare	0.6431544
Entropy RSquare	0.4262358
RMSE	0.4145457
Mean Abs Dev	0.3399483
Misclassification Rate	0.2093023
-LogLikelihood	22.358216
Sum Freq	43

Confusion Matrix

Actual	Predicted Count		
CDR	0	0.5	1
0	23	4	0
0.5	0	8	2
1	0	3	3

Confusion Rates

Actual	Predicted Rate		
CDR	0	0.5	1
0	0.852	0.148	0.000
0.5	0.000	0.800	0.200
1	0.000	0.500	0.500

Test

CDR

Measures	Value
Generalized RSquare	0.4827852
Entropy RSquare	0.2831693
RMSE	0.4792829
Mean Abs Dev	0.3841756
Misclassification Rate	0.3571429
-LogLikelihood	27.701319
Sum Freq	42

Confusion Matrix

Actual	Predicted Count		
CDR	0	0.5	1
0	23	1	1
0.5	7	3	2
1	0	4	1

Confusion Rates

Actual	Predicted Rate		
CDR	0	0.5	1
0	0.920	0.040	0.040
0.5	0.583	0.250	0.167
1	0.000	0.800	0.200

Data Tools & Techniques - *Nominal Logistic (Classification)*

- Misclassification Rate: 24.29%
- Significant predictor variables (p<0.05):
  - MMSE
  - nWBV
- Same for if run for forward/backward analysis

Nominal Logistic Fit for CDR

Effect Summary

Source	LogWorth	PValue
MMSE	16.115	0.00000
nWBV	2.541	0.00288

Remove Add Edit Undo FDR

Converged in Gradient, 6 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	54.89309	4	109.7862	<.0001*
Full	75.83972			
Reduced	130.73281			

RSquare (U)	0.4199
AICc	164.311
BIC	181.329
Observations (or Sum Wgts)	140

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4199	0.4305	0.2404	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6428	0.6552	0.4241	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5417	0.5364	0.6958	$\sum$ -Log(ρ[j])/n
RMSE	0.4214	0.4086	0.4966	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.3290	0.3229	0.3880	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.2429	0.2553	0.3913	$\sum$ (ρ[j]≠ρMax)/n
N	140	47	46	n

Confusion Matrix

Training

Actual	Predicted Count		
CDR	0	0.5	1
0	73	8	0
0.5	15	23	4
1	1	6	10

Validation

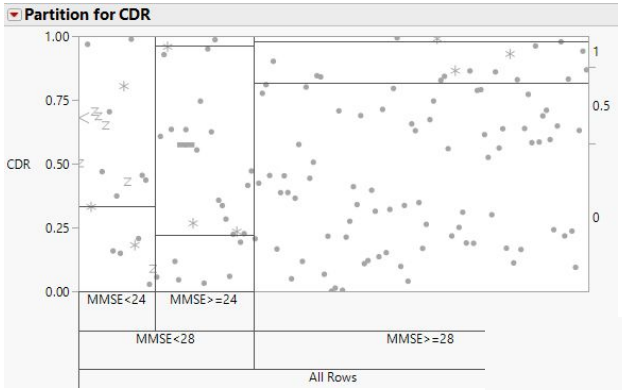
Actual	Predicted Count		
CDR	0	0.5	1
0	23	4	0
0.5	2	9	3
1	0	3	3

Test

Actual	Predicted Count		
CDR	0	0.5	1
0	23	4	0
0.5	7	3	4
1	0	3	2

Data Tools & Techniques – *Decision Tree (Classification)*

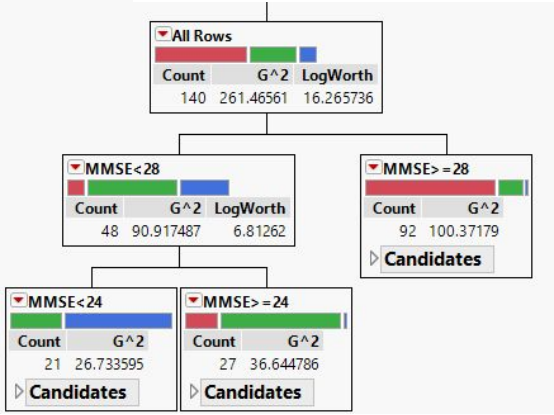
- Number of Splits: 2
- Partition Variables:
  - MMSE (<24 or >=24)
  - MMSE (<28 or >=28)
- Misclassification rate: 22.14%



	RSquare	N	Number of Splits
Training	0.370	140	2
Validation	0.377	47	
Test	0.131	46	

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3696	0.3768	0.1308	1-Loglike(mod
Generalized RSquare	0.5897	0.5994	0.2537	(1-(L(0)/L(mod
Mean -Log p	0.5887	0.5870	0.7962	∑ -Log(ρ[j])/n
RMSE	0.4298	0.4360	0.5204	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3518	0.3623	0.4328	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2214	0.2553	0.3913	∑ (ρ[j]≠ρMax)/
N	140	47	46	n



Confusion Matrix

Training

Actual	Predicted Count		
CDR	0	0.5	1
0	75	6	0
0.5	15	20	7
1	2	1	14

Validation

Actual	Predicted Count		
CDR	0	0.5	1
0	24	3	0
0.5	3	6	5
1	0	1	5

Test

Actual	Predicted Count		
CDR	0	0.5	1
0	24	3	0
0.5	7	1	6
1	0	2	3

Data Tools & Techniques – *Bootstrap Forrest (Classification)*

- Misclassification Rate: 30%
- Number of Trees: 4

Bootstrap Forest for CDR

Specifications

Target Column:	CDR	Training Rows:	140
Validation Column:	Validation	Validation Rows:	47
		Test Rows:	46
Number of Trees in the Forest:	4	Number of Terms:	7
Number of Terms Sampled per Split:	1	Bootstrap Samples:	140
		Minimum Splits per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3038	0.2134	0.1948	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5121	0.3904	0.3573	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6501	0.7409	0.7376	$\sum$ -Log(ρ[j])/n
RMSE	0.4805	0.5147	0.5188	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.4328	0.4729	0.4774	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.3000	0.3191	0.3478	$\sum$ (ρ[j]≠ρMax)/n
N	140	47	46	n

Confusion Matrix

Training

Actual	Predicted Count		
CDR	0	0.5	1
0	78	3	0
0.5	26	15	1
1	4	8	5

Validation

Actual	Predicted Count		
CDR	0	0.5	1
0	24	3	0
0.5	7	7	0
1	1	4	1

Test

Actual	Predicted Count		
CDR	0	0.5	1
0	26	1	0
0.5	8	4	2
1	1	4	0

Data Tools & Techniques – *Nominal Logistic (Prediction)*

- R-Squared: 24.29%
- Significant predictor variables (p<0.05):
  - MMSE
  - nWBV
  - Age (borderline, left in)
- Same for if run for forward/backward analysis

Nominal Logistic Fit for CDR 0/1

Effect Summary

Source	LogWorth	PValue
MMSE	13.576	0.00000
nWBV	3.136	0.00073
Age	1.049	0.08941

Remove Add Edit Undo ☐ FDR

Converged in Gradient, 6 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	43.913752	3	87.8275	<.0001*
Full	51.391096			
Reduced	95.304848			

RSquare (U)	0.4608
AICc	111.078
BIC	122.549
Observations (or Sum Wgts)	140

Fit Details

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	135	51.391096	102.7822
Saturated	138	0.000000	Prob>ChiSq
Fitted	3	51.391096	0.9822

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-48.458646	10.928451	19.66	<.0001*
Age	0.05625953	0.0333529	2.85	0.0916
MMSE	0.86117585	0.1781766	23.36	<.0001*
nWBV	27.6147279	8.7303461	10.01	0.0016*

For log odds of 0/1

Covariance of Estimates

Confusion Matrix

Training

Actual CDR 0/1	Predicted Count 0	1
0	72	9
1	16	43

Validation

Actual CDR 0/1	Predicted Count 0	1
0	26	1
1	1	18

Test

Actual CDR 0/1	Predicted Count 0	1
0	21	6
1	7	13