<u>OPIM 5604 Project Report</u>

Group 2
Kristine Carnavos
Edgar Lorenzo
Rahul Sinha
Mary Mulrooney

<u>Paper Outline</u>

Alzheimer's Disease/Dementia Classification Model

Team 2:

Kristine Carnavos

Edgar Lorenzo

Rahul Sinha

Mary Mulrooney

OPIM 5604: Predictive Modeling

Project White Paper

Due Date: 4/29/19

Problem Statement/Introduction

In the United States Alzheimer's in the most common form of dementia. It affects over 5.7 million Americans every year. Of all Americans who are 65 or older 10% of them have some form of the disease. There is no known cure for Alzheimer's disease. Currently all that is available are therapies which help lessen the severity of symptoms. These therapies significantly increase a patient's quality of life. The lifetime cost of care for an individual with Alzheimer's in 2018 was $350,174. 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.

A diagnosis can only be achieved through a combination of extensive cognitive testing, blood test, and MRI scans which can be very expensive. To date there is not one set tool which can definitively diagnose Alzheimer's. The only sure way to detect Alzheimer's currently is to observe a patient once symptoms have already started.

With our research we hope to develop an understanding of the possible markers of the disease before symptoms begin to develop. Our aim is to produce a classification model that will be able to diagnose the level of severity of Alzheimer’s to allow for correct therapies & treatments to be prescribed to patients. This will allow for early treatment and intervention on patients who have been identified as having the most critical input variables. This can possibly lessen the severity that the disease has on the patient's life.

Early detection will also aid in reducing the annual cost of Alzheimer's disease for American affected with the disease. This will in turn lower the national cost of Alzheimer's not only in the United States, but worldwide.

Data Set Summary

The data set was obtained from Kaggle.com: *MRI and Alzheimer's: Magnetic Resonance Imaging Comparisons of Demented and Nondemented Adults* by Jacob Boysen. It was taken from the Open Access Series of Imaging Studies (OASIS). This data set is a collection of 416 unique subjects, ages 18 to 96, and included cross-sectional MRI scan data as well as other factors and tests that help to ultimately determine their applicable dementia/Alzheimer's Disease rating. A control group of 20 subjects was included in the data that have no dementia rating. The full details of the data set from Kaggle.com can be found in Appendix A . Some of the variables included in the full data set are meant for subject identification and control. Table 1 shows the predictor variables included in our analysis with the applicable data types and response levels (if nominal).

Table 1: Predictor Variables for Dementia/Alzheimer's Disease Classification Model

Predictor Variable	Data Type	Response Levels (if applicable)
Gender	Nominal	Male, Female
Age	Continuous	N/A
Education Level	Nominal	1 to 5
Socioeconomic Status	Nominal	1 to 5
Mini Mental State Exam (MMSE)	Continuous	N/A
Estimated Total Intracranial Volume (eTIV)	Continuous	N/A
Normalize Whole Brain Volume (nWBV)	Continuous	N/A
Atlas Scaling Factor (ASF)	Continuous	N/A

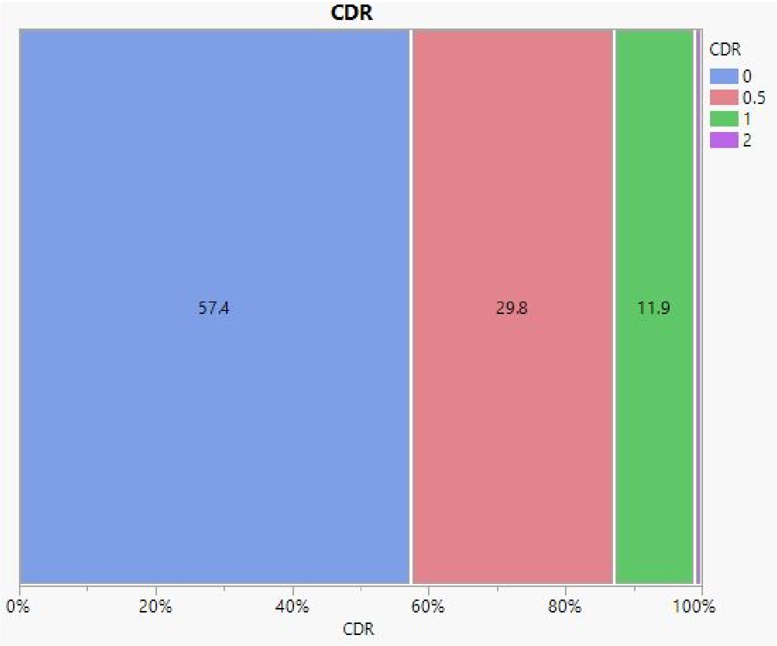
The target variable for this data set is **CDR**, which stands for the **Clinical Dementia Rating**. This is the method that ultimately is used by doctors to determine the severity of the dementia diagnosis. For this data set, the ratings included are as follows:

- 0: No Dementia/Alzheimer's Disease
- 0.5: Very Mild Dementia/Alzheimer's Disease

- 1: Mild Dementia/Alzheimer's Disease
- 2: Moderate Dementia/Alzheimer's Disease

A mosaic plot was created of CDR to understand the distribution of the variable in the data set and see if any one category is over-represented in the data set. The results can be seen in the figure below. As seen, 57.4% of the data set is represented by subjects with a CDR rating of 0, meaning they have no diagnosis of dementia/Alzheimer's Disease. This is a bit concerning, as it may skew the data set towards being a better predictor of those with no dementia/Alzheimer's than being able to classify the levels of dementia/Alzheimers. Also, it is observed that only 2 lines of data included in the set represents subjects with a CDR rating of 2, which is only 0.9% of the data. In order to reduce unnecessary complexity in modeling the data set, these lines have been excluded from the classification model analysis.

Figure 1: Distribution of CDR by Response Level



Goals of Analysis

The goal of the analysis was to determine which variables, if any, had a strong impact to predict our target variable of CDR. Specifically, it was of interest to see if education level and socioeconomic status have a link to developing Alzheimer's disease and what the variation is. It would also be of interest to health care providers to see how the Mini-Mental State Examination (MMSE) is correlated to dementia. This is a 30-point questionnaire that is a quick, simple, and costless exam that all providers can use to test their patients. If this is an indicator of dementia, perhaps we can lobby that all physicians must perform this exam annually for their patients. This could allow for early detection and early treatment with our drug (and more patients we are able to treat for a longer period of time). From there, classification models were created with those outputs, and the misclassification rates were used to determine which model should be selected for our purposes. Ideally, we wanted to minimize our misclassification rate, with a goal of <5% if possible.

An alternative approach was also later considered for the data set once the distribution of CDR ratings in the data set was observed. In this approach, the CDR rating was converted to a 0/1 rating where the original level of 0 stands for “Subject does not have dementia/Alzheimer’s disease” as in the classification model, and 1 is made up of the original 0.5, 1, and 2-rated subjects and stands for “Subjects have Dementia/Alzheimer’s Disease.” In the analysis, it was of interest if the same variables would remain significant as seen in the classification model as well as what amount of variation was explained by the predictor models and how the misclassification rates of the selected predictor model would compare to the selected classification model.

Even though the predictor model would be of less use in determining specific therapies and drugs for early intervention, it would still allow physicians to have a level of understanding of which subjects are likely to get dementia/Alzheimer's disease and start to work with them to determine which therapies/drugs work best for patients.

Data Analysis

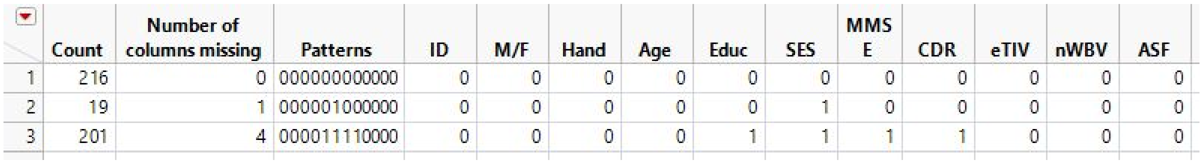
Data Cleaning

The data set was first analyzed to understand if there was a significant amount of missing data. The full details of the missing data analysis can be seen below in figure 2. The following observations were noted from the missing data analysis:

- 201 of the 436 subjects were missing both the target variable, CDR, as well as predictor variables SES, Education, and MMSE
- SES was noted to be missing 19 lines of data from the remaining 235 lines of data, encompassing 8% of the total data set.

For the first, the missing data encompassed 49.5% of the data set, a significant amount of missing data. Since the target variable as well as multiple predictor variables are missing, these rows of data were excluded from the analysis. For the second, predictor variable SES is noted to be of data type nominal. Because of this, imputing is not possible. In this case, the missing variables were ignored and the 19 lines of data continued to be included in the data set used for the analysis.

Figure 2: Missing Data Analysis Results



	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	201	4	000011110000	0	0	0	0	1	1	1	1	0	0	0

Outliers Analysis

Each of the predictor variables were also plotted in order to determine if any outliers exist. Both MMSE and eTIV were noted to have outliers, but these values comprised less than 5% of the data set. So for these variables, the outliers were ignored. Distribution of the data and outliers can be seen in figures 3 and 4 below.

Figure 3: MMSE Data Set Distribution w/ Outliers Indicated

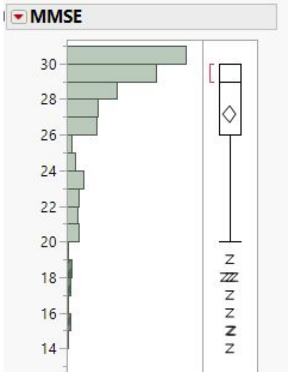
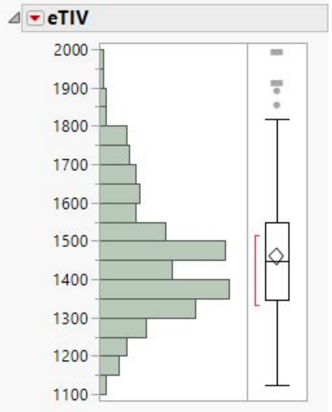


Figure 4: eTIV Data Set Distribution w/ Outliers Indicated



Graphical Summary & Hypothesis Testing

The data was visualized via various charts and graphs to better understand the predictor variables and their effects on CDR. First, mosaic plots were created for predictor variables **Gender, SES,** and **eTIV**. A contingency analysis was conducted to understand if there is any noted difference between the predictor variable levels and their CDR level. For these variables, p-values of 0.1361, 0.2522, and 0.2540 were observed, indicating that there is no difference between these groups' CDR level. This indicates that these variables will not be expected to be significant predictors of CDR.

Figure 5: Mosaic Plot of CDR by Gender (M/F)

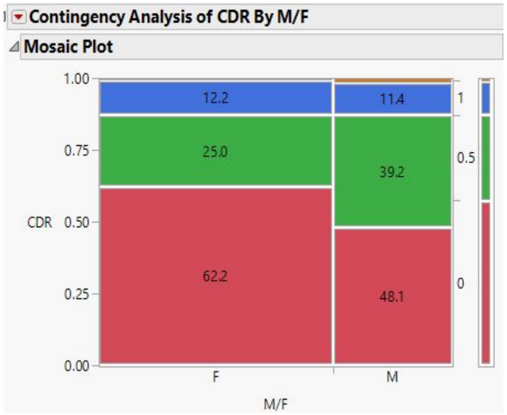


Figure 6: Mosaic Plot of CDR by SES (1-5)

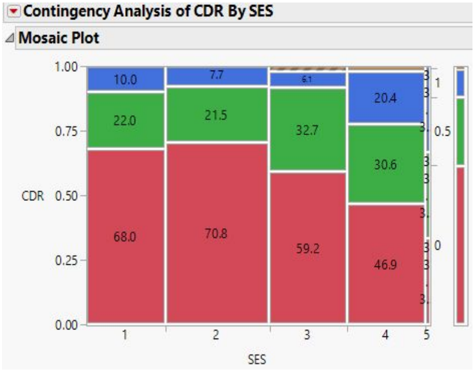
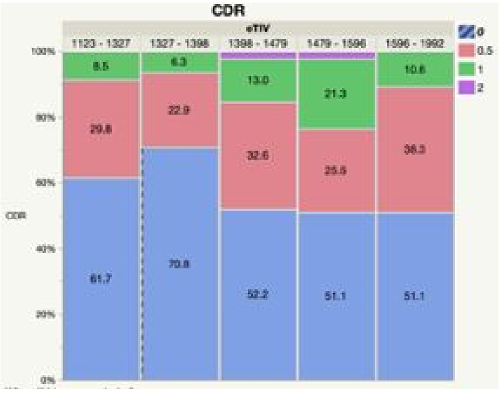


Figure 7: Mosaic Plot of CDR by eTIV



Mosaic plots of variables **Age, nWBV, and MMSE** were also created, and each of the corresponding contingency analyses show that each of these variables are shown to be different when sorted by CDR with p-values of <0.0001. Ultimately, it is expected that when classification models are created, these three variables will remain significant predictors of CDR. The mosaic plots of each of the noted significant variables can be found in figures 8, 9, and 10 below.

Other observations noted from this visualization is that there seems to be a negative correlation between **nWBV** and **Age.** Also**,** greater educational attainment was associated with a lower likelihood of dementia due to Alzheimer's.

Figure 8: Mosaic Plot of CDR by Age

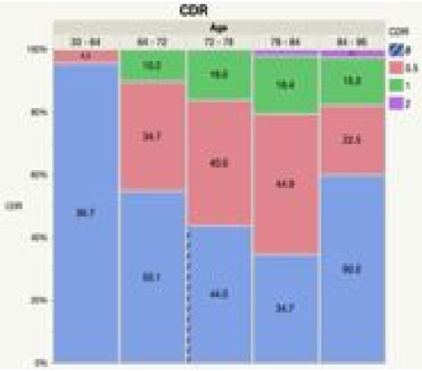


Figure 9: Mosaic Plot of CDR by nWBV

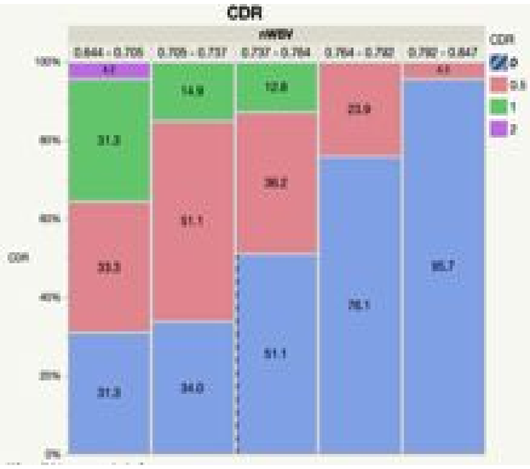
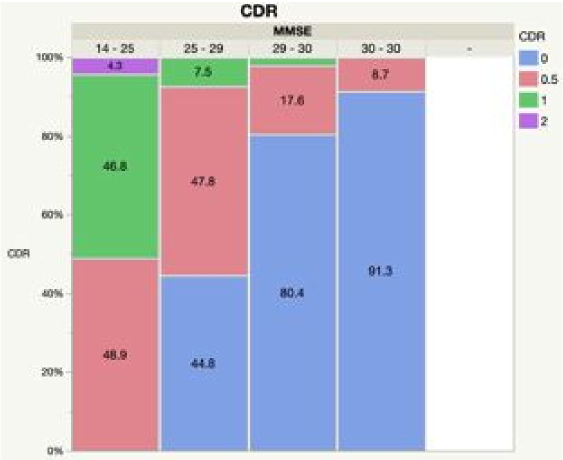


Figure 10: Mosaic Plot of CDR by nWBV



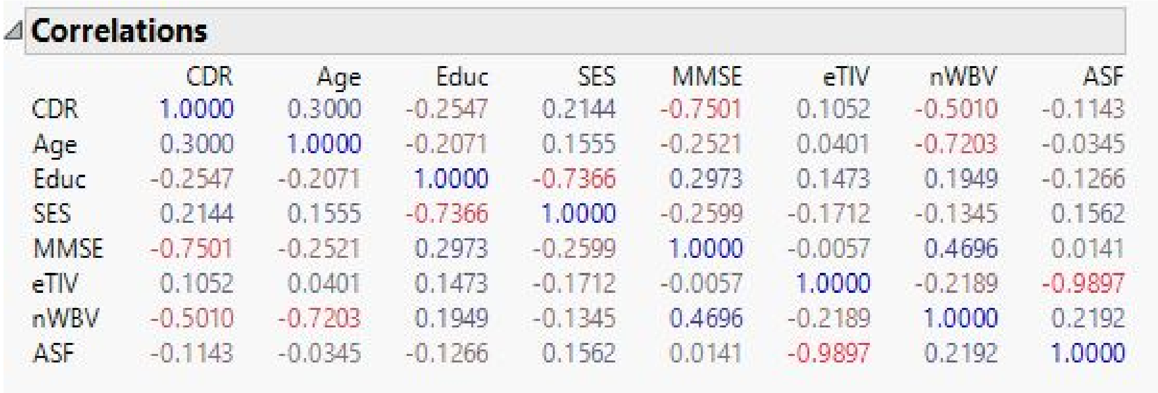
Correlation Matrix & Principal Components Analysis

A correlations analysis was conducted for all variables to understand which are shown to be of the greatest significance to predicting CDR as well as if any are correlated to each other. The correlation matrix is shown in figure 11 below. From this analysis, **MMSE** and **nWBV** are noted to have high negative correlations to **CDR** (correlation coefficients of -0.7501 and -0.5010 respectively). It is expected that they will remain significant for all classification models created.

For predictor variable correlations, **ASF and eTIV** were found to be highly negatively correlated at 98.97% and were not strongly correlated to any other variables. Because of this, a principal components analysis was conducted for these two variables, creating a new variable **Princ1**. The Princ1 variable was used in all models created in place of the eTIV and ASF variables. For the number of components, in using both the eigenvalue and percent of variation explained methods, the number of components chosen was 1 (eigenvalue for 1 component is 1.9897, percent of variation explained 99.487%). The resultant formula created for the new **Princ1** variable is:

Princ1= 0.0044092741*eTIV + -5.463358721*ASF + 0.210907798

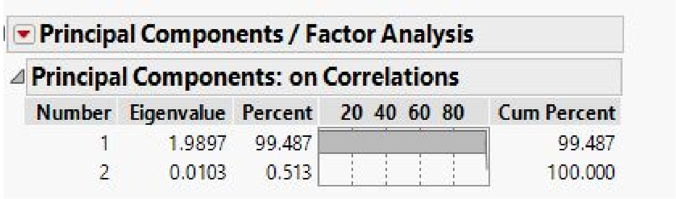
Figure 11: Correlations Matrix for Data Set



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

Figure 12: PCA Analysis for eTIV and ASF



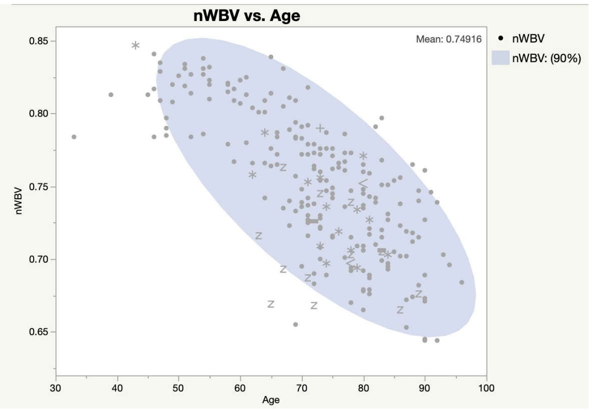
Principal Components / Factor Analysis

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

Other strong negative correlations were 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. The negative correlation is evident in figure 13 below:

Figure 13: nWBV vs. Age



Model Summary & Comparison

Validation Column

For both model types, a validation column was created. For the purposes of this model type, we chose to create a 60/20/20 split between the training, validation, and test set. Although the data set had been significantly truncated due to the missing data, we chose to still include the test set due to the nature of the data. Medical data is very sensitive, and incorrect classification of data could lead to serious repercussions for the medical provider. The target variable type is nominal, so the data was stratified random across this variable to form the validation column.

Classification Models

The models considered for the classification model are as follows:

1. Nominal Logistic Regression with p-values less than 0.05 for all variables
2. Partition/Decision Tree
3. Bootstrap Forest
4. Neural Network

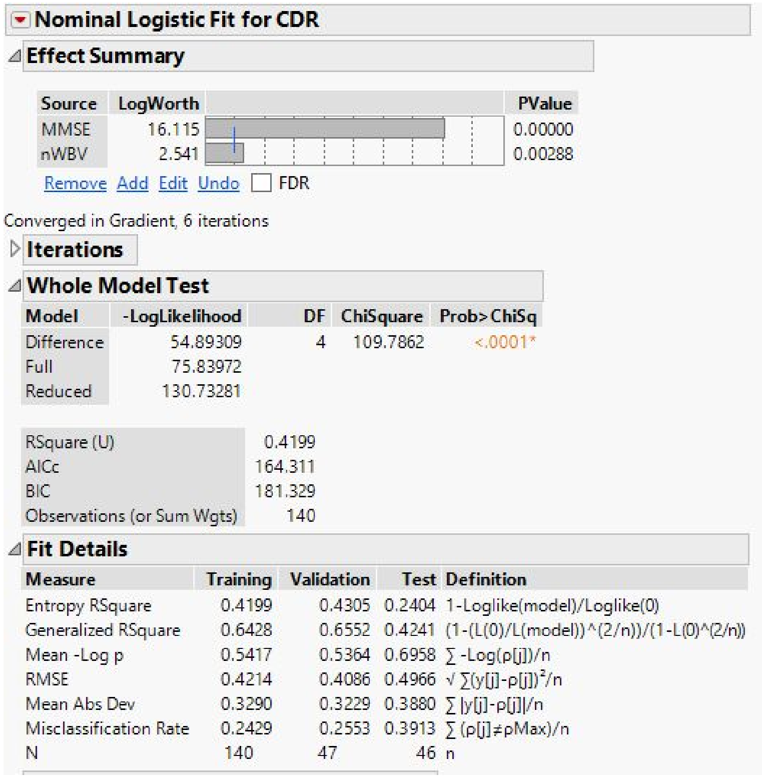
A boosted tree model was not considered for the classification model because there were three levels of classification, which did not comply with the model settings. This was later considered in the prediction model.

Nominal Logistic Regression- Classification

The nominal logistic regression model was formed initially using all of the variables, and then eliminated by using the p-value to evaluate their impact on the model. Ultimately, only two variables remained significant, **MMSE** with a p-value of 0.000 and **nWBV** with a p-value of 0.00288. This model was run both forwards, backwards, and stepwise, and the significant variables remained the same for all versions of the model run. So, the original forward model was used. The detailed results of the model can be seen in Figure 14.

The **misclassification rate** for this model was noted to be **25.53%** for the validation column. This is only slightly greater than that of the test, at 24.29%. This is slightly higher than desired for the type of medial data being used, so the confusion matrices were also analyzed. For the validation confusion matrix, some false positives are noted for the 0/0.5 levels and 0.5/1 levels. This is not ideal for medical data, as it could result in unnecessary treatment for patients. On the other hand, a few false negatives are also noted for 0/0.5 and 0.5/1 levels, which is also not ideal as it will result in patients not being diagnosed, missing out on treatments that could be used to lessen the progress of the disease. Ultimately, the ideal model would minimize both false positives and false negatives for the model.

Figure 14: Nominal Logistic Fit Model & Confusion Matrix Results for CDR Classification



Nominal Logistic Fit for CDR

Effect Summary

Source	LogWorth	PValue
MMSE	16.115	0.00000
nWBV	2.541	0.00288

Converged in Gradient, 6 iterations

Iterations

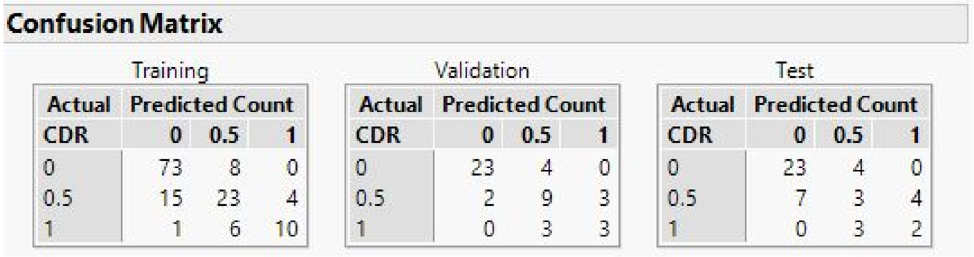
Whole Model Test

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	$\sum$ |y[j]-ρ[j]|/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

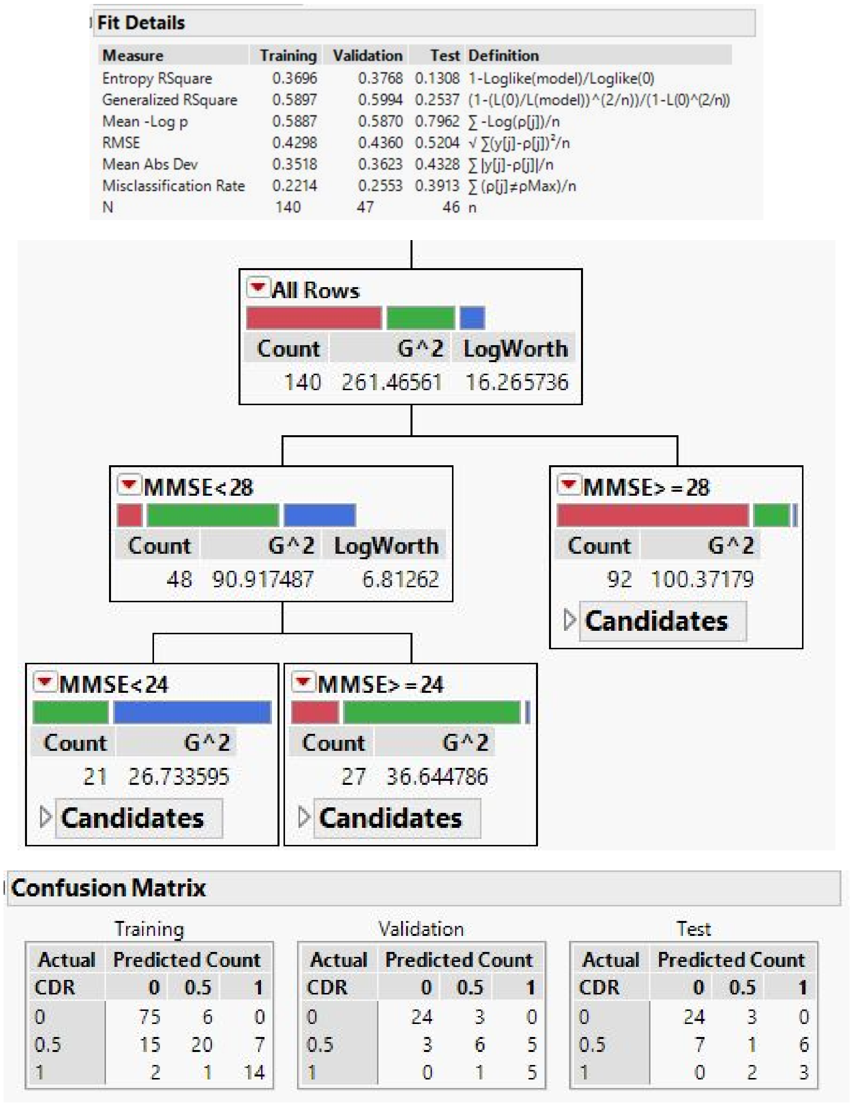
Test

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

Decision Tree/Partition- Classification

For the decision tree/partition model, only one variable remained significant once partitioned, **MMSE**. In this case, the model was partitioned only twice, once with **MMSE</>= 28**, and then again on **MMSE on </>= 24** on the <28 partition. For this model, the misclassification rate was noted to be 25.53% for the validation data set. The detailed results of the model can be seen in figure 15 below. The confusion matrices again show some risk for false positives and negatives in the 0/0.5 and 0.5/1 levels as seen previously in the nominal logistic fit model, but with less false negatives and more false positives. This could be even worse as false positives could result in lawsuits for unnecessary treatments being provided to

Figure 15: Decision Tree/Partition Model & Confusion Matrix Results for CDR Classification



Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.3696	0.3768	0.1308	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5897	0.5994	0.2537	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
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
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

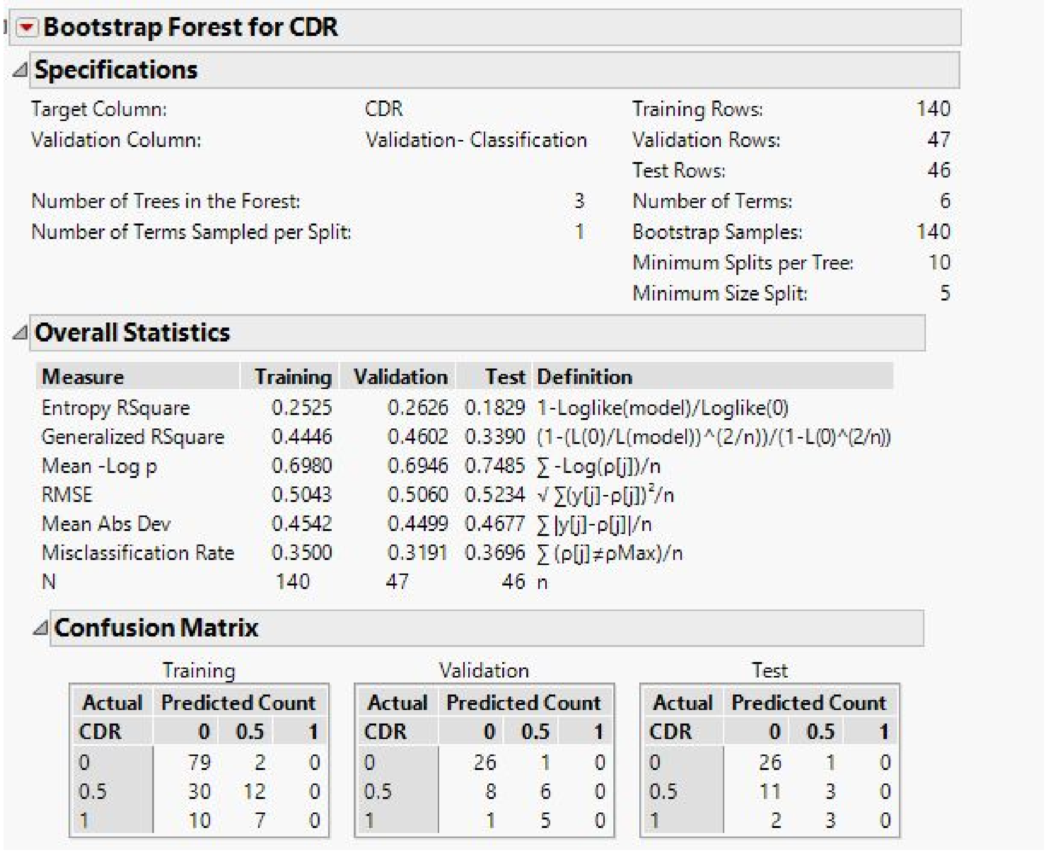
Bootstrap Forest- Classification

For the bootstrap forest model, a total of 3 trees were created from the data. These trees included the following variables:

- Tree 1: Education, MMSE, Age
- Tree 2: Education, Age
- Tree 3: MMSE, Education, nWBV

The misclassification rate for the validation set was noted to be 31.91%, much higher than the last two models. In looking at the confusion matrices, there is also a higher risk of false negatives and positives with this model. It is not expected that this model will be selected for the classification data. Full details on the model results can be seen in Figure 16.

Figure 15: Bootstrap Model & Confusion Matrix Results for CDR Classification



Bootstrap Forest for CDR

Specifications

Target Column:	CDR	Training Rows:	140
Validation Column:	Validation- Classification	Validation Rows:	47
		Test Rows:	46
Number of Trees in the Forest:	3	Number of Terms:	6
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	0.2525	0.2626	0.1829	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4446	0.4602	0.3390	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6980	0.6946	0.7485	∑ -Log(ρ[j])/n
RMSE	0.5043	0.5060	0.5234	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.4542	0.4499	0.4677	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.3500	0.3191	0.3696	∑ (ρ[j]≠ρMax)/n
N	140	47	46	n

Confusion Matrix

Training

Actual	Predicted Count		
CDR	0	0.5	1
0	79	2	0
0.5	30	12	0
1	10	7	0

Validation

Actual	Predicted Count		
CDR	0	0.5	1
0	26	1	0
0.5	8	6	0
1	1	5	0

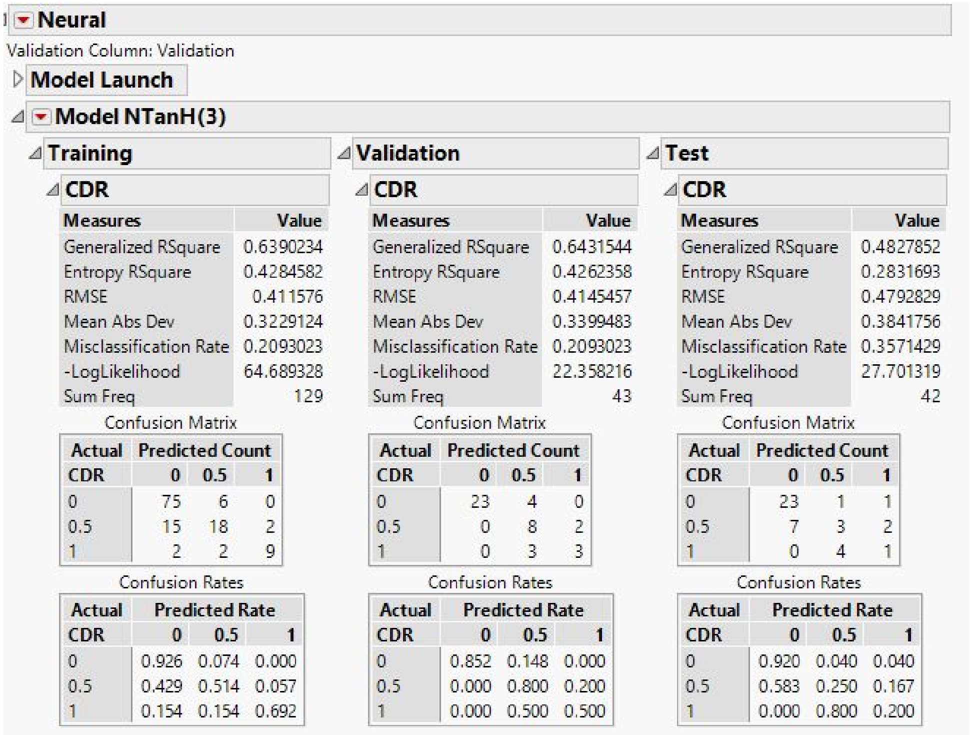
Test

Actual	Predicted Count		
CDR	0	0.5	1
0	26	1	0
0.5	11	3	0
1	2	3	0

Neural Network- Classification

For the neural network, only one level was considered due to the lack of complexity of the data set. All variables were considered in the Neural Network model. For the validation set, the misclassification rate was noted to be 20.93%. There is a small chance of some false positives and negatives noted for the 0/1 and 0.5/1 levels. Full details of the model results can be seen in figure 16 below.

Figure 16: Neural Network Model & Confusion Matrix Results for CDR Classification



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

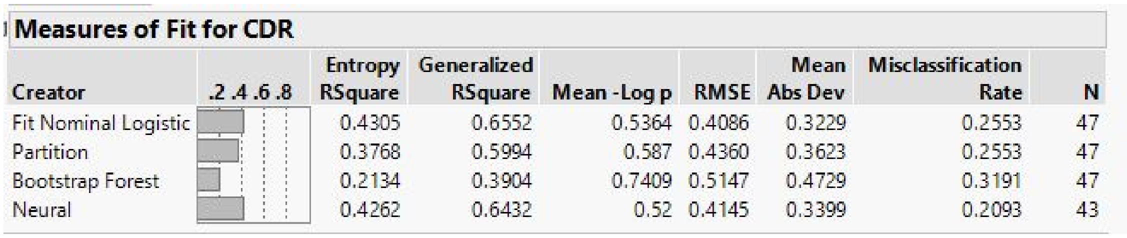
Actual	Predicted Rate		
CDR	0	0.5	1
0	0.920	0.040	0.040
0.5	0.583	0.250	0.167
1	0.000	0.800	0.200

Classification Model Compare/Selection

A model compare was created for all 4 models for the validation set to understand which is the best option for our data. In this case, the model with the lowest misclassification rate is the **Neural Network,** with a **misclassification rate of 20.93%**. This is our selected model.

The misclassification rate of the selected model (and all models considered) is noted to be much higher than desired (<5% for this study, <1% for the medical field), with a great chance for false positives and negatives. In the medical field, we want to try to minimize the amount of these seen to lessen the possibility of medical malpractice suits. The reduction of data as well as CDR distribution (57.4% being of 0 classification) may be attributed to this inflated misclassification rate.

Figure 17: Model Compare for CDR Classification Model (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

Prediction Model

Prediction Model Data Set

As previously noted, the data set was also considered as a prediction model. In this analysis, CDR was re-coded to a 0/1 response with:

- 0: No Dementia/Alzheimer's Disease diagnosis
- 1: Has Dementia/Alzheimer's Disease diagnosis

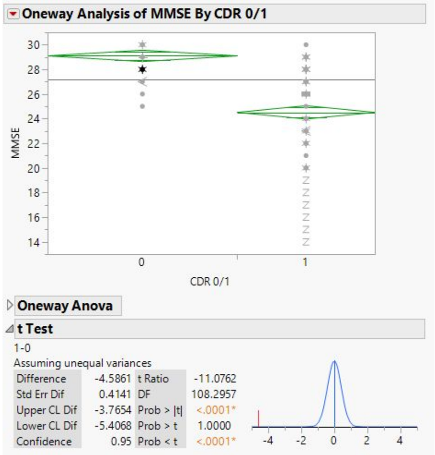
For this analysis, the previously excluded 2 lines of data with a “2: Moderate Dementia/Alzheimer’s Disease” rating were included as a 1 level.

By modifying the data set to this 0/1 response, it allows the target variable to be better distributed in the data set as well as allows us to consider the boosted tree model. The data set now comprises 57.4%/42.6% for 0/1 CDR responses, much better distribution than seen in the classification data set.

Graphical Summary & Hypothesis Testing

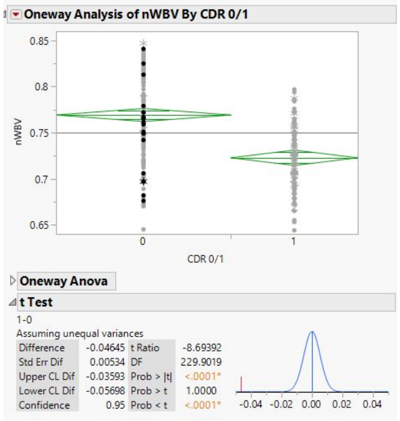
Prior to creating prediction models, the data was visualized to understand if the change to 0/1 CDR response would affect which variables are predicted to be significant for CDR. The predictor variables noted to be statistically different when sorted by CDR 0/1 are MMSE, nWBV, Age, Gender, and eTIV (p<0.0001). Of these 5 variables, Gender and eTIV were not noted to be different in the classification model. This indicates that they may end up being statistically significant in the classification models. The graphical summaries of each of these variables can be seen in figures 18-21 below.

Figure 18: CDR (0/1) by MMSE



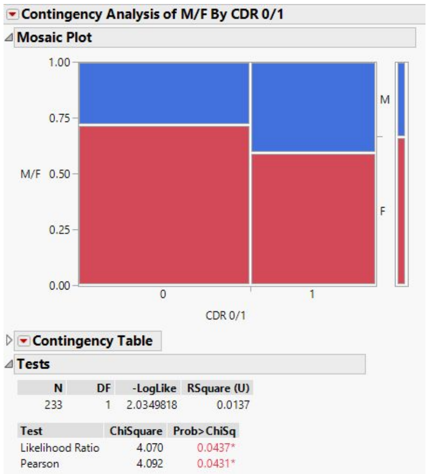
```
1-0
Assuming unequal variances
Difference     -4.5861  t Ratio    -11.0762
Std Err Dif     0.4141  DF         108.2957
Upper CL Dif   -3.7654  Prob > |t|  <.0001*
Lower CL Dif   -5.4068  Prob > t    1.0000
Confidence        0.95  Prob < t    <.0001*
```

Figure 19: CDR (0/1) by nWBV



```
t Test
1-0
Assuming unequal variances
Difference     -0.04645  t Ratio     -8.69392
Std Err Dif     0.00534  DF          229.9019
Upper CL Dif   -0.03593  Prob > |t|   <.0001*
Lower CL Dif   -0.05698  Prob > t      1.0000
Confidence         0.95  Prob < t     <.0001*
```

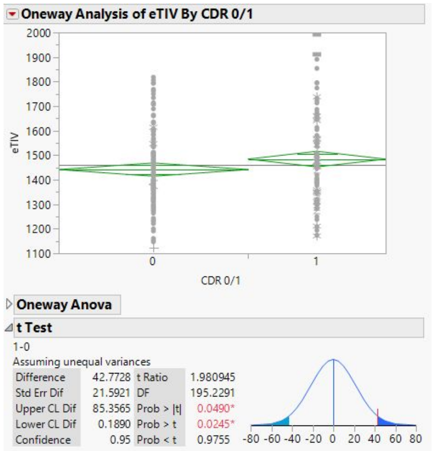
Figure 20: CDR (0/1) by Gender



```
Tests
  N   DF   -LogLike   RSquare (U)
233    1  2.0349818        0.0137

Test              ChiSquare  Prob>ChiSq
Likelihood Ratio      4.070     0.0437*
Pearson               4.092     0.0431*
```

Figure 21: CDR (0/1) by eTIV



```
t Test
1-0
Assuming unequal variances
Difference    42.7728  t Ratio    1.980945
Std Err Dif   21.5921  DF         195.2291
Upper CL Dif  85.3565  Prob > |t|  0.0490*
Lower CL Dif   0.1890  Prob > t    0.0245*
Confidence       0.95  Prob < t     0.9755
```

Validation Column

A new validation column was created for the prediction model data set, but the 60/20/20 split between the training, validation, and test set was maintained. Data was stratified random on the re-coded 0/1 CDR variable.

Prediction Models Considered

The models considered for the prediction model are as follows:

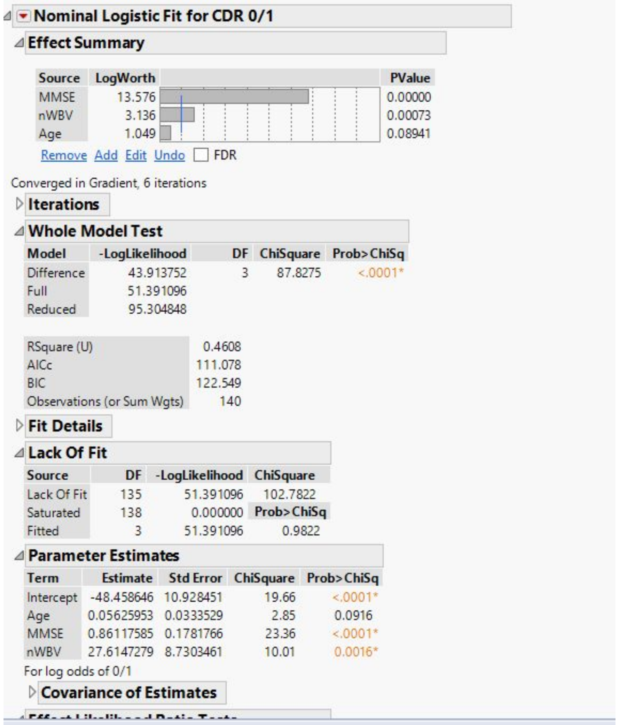
1. Nominal Logistic Regression with p-values less than 0.05 for all variables
2. Decision Tree/Partition
3. Bootstrap Forest
4. Boosted Tree
5. Neural Network

Nominal Logistic Regression- Prediction

The nominal logistic regression model was formed initially using all of the variables, and then eliminated by using the p-value to evaluate their impact on the model. Ultimately, only two variables remained significant, **MMSE** with a p-value of 0.000 and **nWBV** with a p-value of 0.00073. Age was also left in the model, with a p-value of 0.08941 as it was border-line, and it did not increase the complexity of the model much. This model was run both forwards, backwards, and stepwise, and the significant variables remained the same for all versions of the model run. The detailed results of the model can be seen in Figure 22.

The R-Squared value for the validation was found to be 75.17%, and the misclassification rate was minimized to be 4.26%. Looking at the confusion matrices, the validation set seems to have minimal occurrences of false positives and negatives. Looking at the training and test values, however, there is quite a discrepancy between the values seen.

Figure 22: Nominal Logistic Fit Model & Confusion Matrix Results for CDR- Prediction



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

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

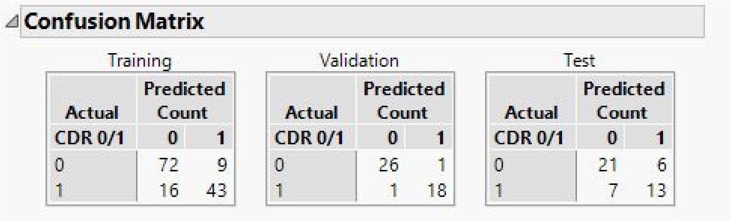
Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	135	51.391096	102.7822
Saturated	138	0.000000	Prob>ChiSq
Fitted	3	51.391096	0.9822

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

Validation

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

Test

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

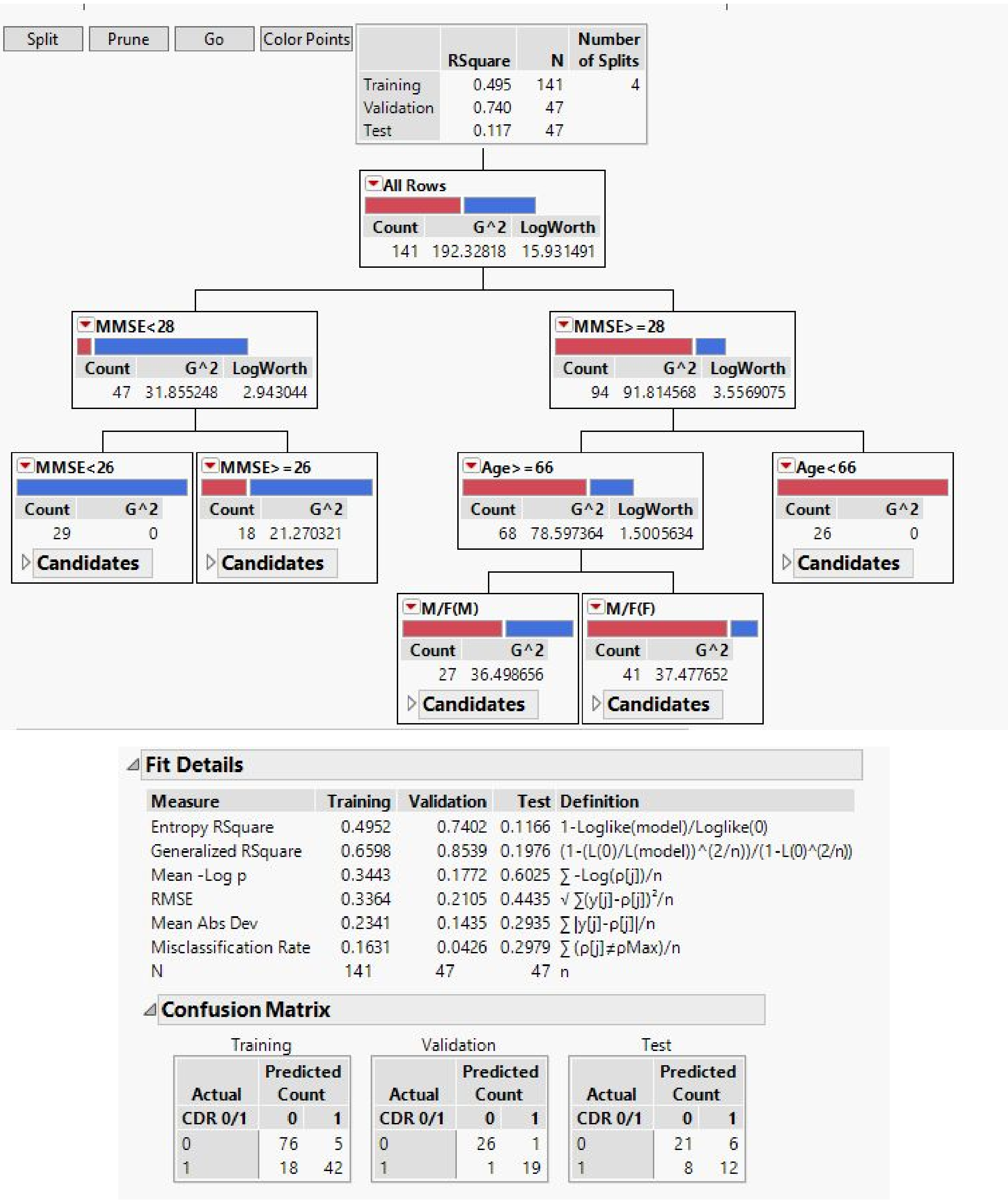
Decision Tree/Partition- Prediction

For the decision tree, 4 splits were created for the prediction model data set. The relevant variables included are:

- MMSE
- Age
- Gender

The first split mimicked the first split seen in the classification data set, with MMSE</>= 28, but it diverges from there. Under MMSE<28, it then splits MMSE </>=26, which is different than the split on 24 from the classification model. Under the MMSE>=28 column, it splits by Age >=/< 66. Finally, under Age>=66, it splits by gender. Overall, the R-Squared Value is seen to be 74.0% on the Validation set, and the misclassification rate is 4.26%. The confusion matrix for the validation data set shows again a very small chance of false positives/negatives. The same discrepancy between the training and test and validation data sets are noted as seen in the nominal fit regression results. Full model detail results can be seen in Figure 23.

Figure 23: Decision Tree/Partition Model & Confusion Matrix Results for CDR- Prediction



	RSquare	N	Number of Splits
Training	0.495	141	4
Validation	0.740	47	
Test	0.117	47	

All Rows

Count	G^2	LogWorth
141	192.32818	15.931491

MMSE<28

Count	G^2	LogWorth
47	31.855248	2.943044

MMSE>=28

Count	G^2	LogWorth
94	91.814568	3.5569075

MMSE<26

Count	G^2
29	0

Candidates

MMSE>=26

Count	G^2
18	21.270321

Candidates

Age>=66

Count	G^2	LogWorth
68	78.597364	1.5005634

Age<66

Count	G^2
26	0

Candidates

M/F(M)

Count	G^2
27	36.498656

Candidates

M/F(F)

Count	G^2
41	37.477652

Candidates

Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.4952	0.7402	0.1166	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6598	0.8539	0.1976	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.3443	0.1772	0.6025	∑ -Log(p[j])/n
RMSE	0.3364	0.2105	0.4435	√ ∑(y[j]-p[j])²/n
Mean Abs Dev	0.2341	0.1435	0.2935	∑ |y[j]-p[j]|/n
Misclassification Rate	0.1631	0.0426	0.2979	∑ (p[j]≠pMax)/n
N	141	47	47	n

Confusion Matrix

Training

Actual	Predicted Count	
CDR 0/1	0	1
0	76	5
1	18	42

Validation

Actual	Predicted Count	
CDR 0/1	0	1
0	26	1
1	1	19

Test

Actual	Predicted Count	
CDR 0/1	0	1
0	21	6
1	8	12

Bootstrap Forest- Prediction

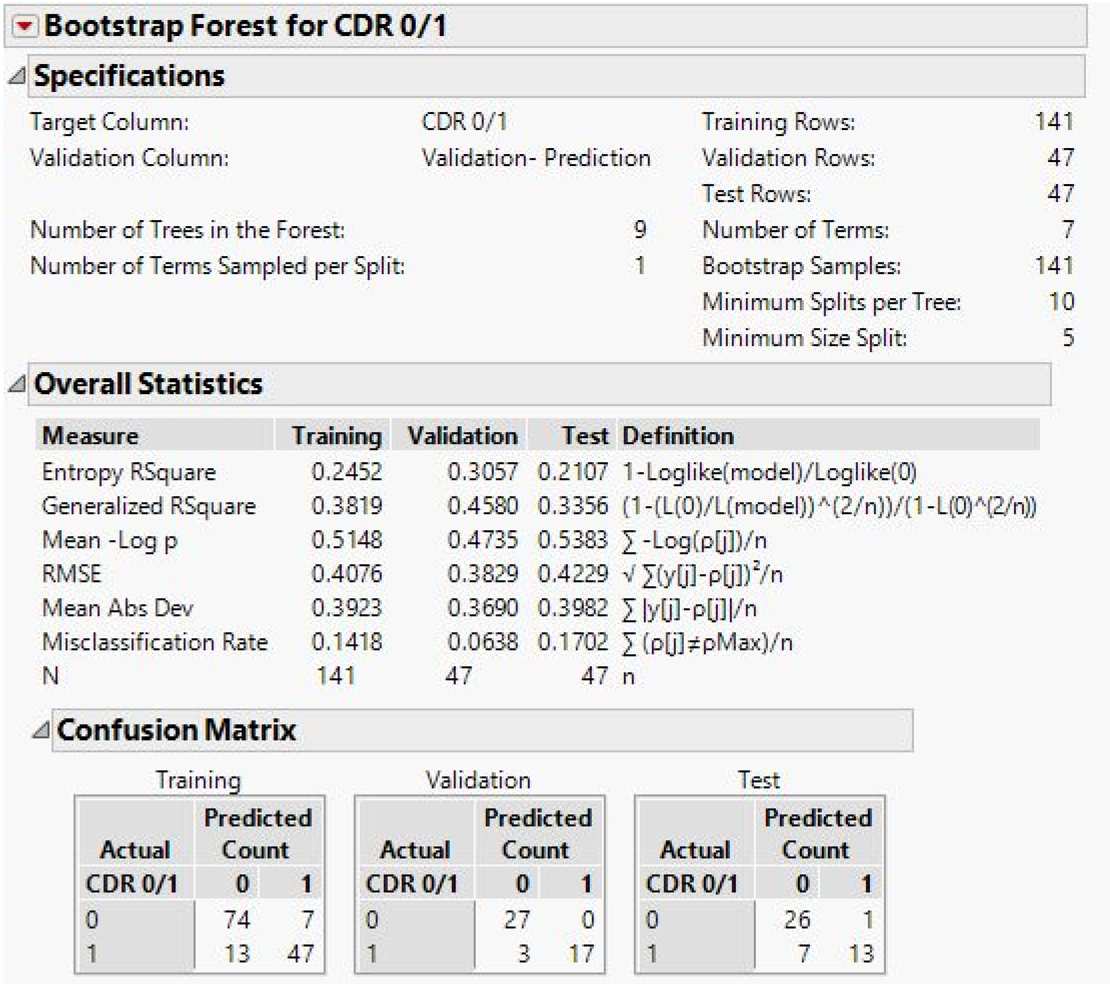
The bootstrap forest model for the prediction data set yielded 9 trees in the forest. The trees involve the following variables:

- Tree 1: Education, Gender
- Tree 2: nWBV
- Tree 3: MMSE, Education, Princ1
- Tree 4: nWBV
- Tree 5: Princ1, SES, Gender

- Tree 6: Princ1, Education, nWBV
- Tree 7: MMSE, SES
- Tree 8: Princ1, Age
- Tree 9: Gender, Age, Education, nWBV

This model was noted to only have an R-Squared value of 30.57%, but the misclassification rate was noted to still be fairly low at 6.38%. The confusion matrix shows a risk for false negatives in the validation set, which is less of a risk for lawsuits but will result in no treatments being prescribed to those that need it. Ultimately, this is not an idea model for our purposes. Full details of model results can be seen in Figure 24.

Figure 24: Bootstrap Forest Model & Confusion Matrix Results for CDR- Prediction



Bootstrap Forest for CDR 0/1

Specifications

Target Column:	CDR 0/1	Training Rows:	141
Validation Column:	Validation- Prediction	Validation Rows:	47
		Test Rows:	47
Number of Trees in the Forest:	9	Number of Terms:	7
Number of Terms Sampled per Split:	1	Bootstrap Samples:	141
		Minimum Splits per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2452	0.3057	0.2107	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3819	0.4580	0.3356	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5148	0.4735	0.5383	∑ -Log(ρ[j])/n
RMSE	0.4076	0.3829	0.4229	√ ∑(y[j]-ρ[j])^2/n
Mean Abs Dev	0.3923	0.3690	0.3982	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1418	0.0638	0.1702	∑ (ρ[j]≠ρMax)/n
N	141	47	47	n

Confusion Matrix

Training

Actual	Predicted Count	
CDR 0/1	0	1
0	74	7
1	13	47

Validation

Actual	Predicted Count	
CDR 0/1	0	1
0	27	0
1	3	17

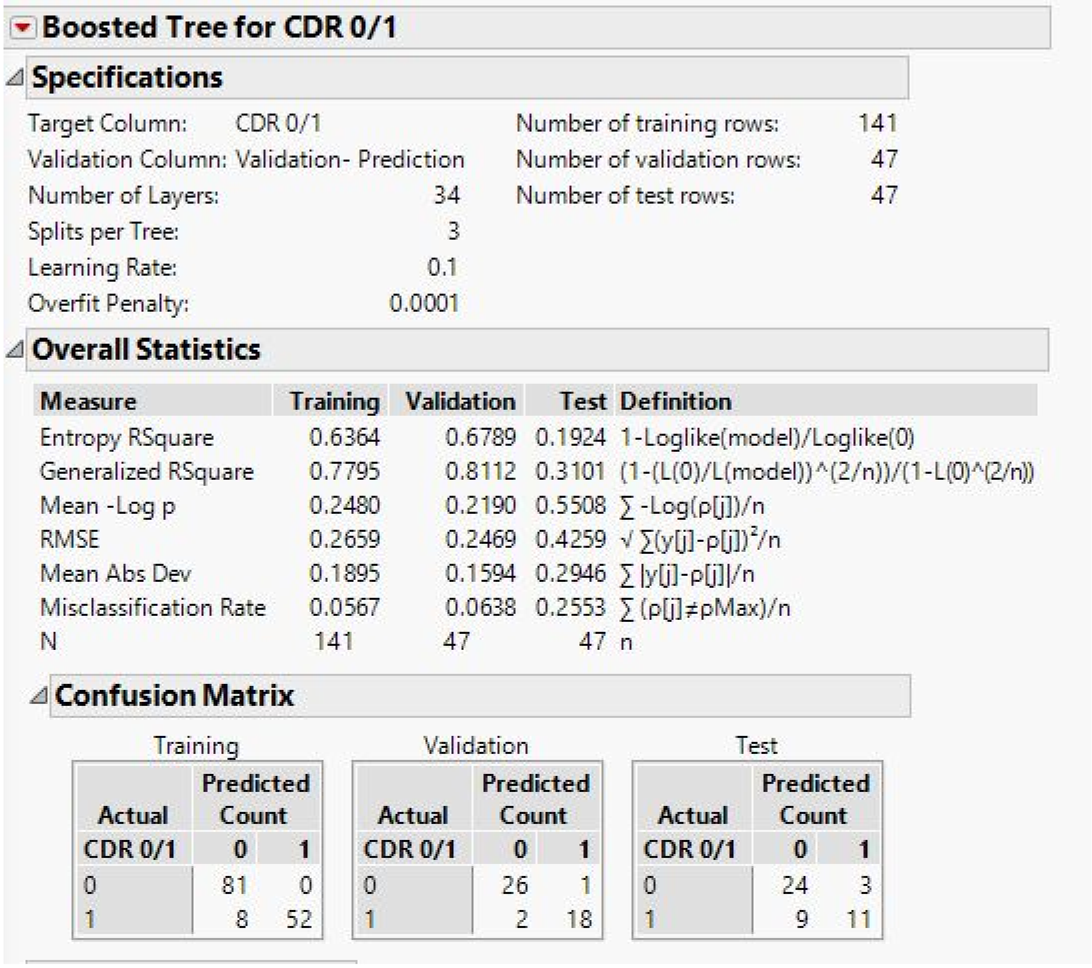
Test

Actual	Predicted Count	
CDR 0/1	0	1
0	26	1
1	7	13

Boosted Tree- Prediction

Due to the reduction in complexity of the model target variable, a boosted tree was able to be run for our prediction data set. For the boosted tree, the learning rate was set to 10% to reduce risk to model. Overall, the number of layers used for our data set was 34, with 3 splits per tree. All variables were considered for this model. The R-Squared value of the validation set was 67.89%, a bit lower than some of the others. The misclassification rate was noted to be 6.38%, which is higher than expected considering the R-Squared value. Looking at the confusion matrix, there is a risk of some false positives and negatives with this model, which is not ideal and may explain the higher misclassification rate. It is likely that this model will not be selected for use. Full details of model results can be seen in Figure 25.

Figure 25: Boosted Tree Model & Confusion Matrix Results for CDR- Prediction



Boosted Tree for CDR 0/1

Specifications

Target Column:	CDR 0/1	Number of training rows:	141
Validation Column:	Validation- Prediction	Number of validation rows:	47
Number of Layers:	34	Number of test rows:	47
Splits per Tree:	3		
Learning Rate:	0.1		
Overfit Penalty:	0.0001		

Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.6364	0.6789	0.1924	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7795	0.8112	0.3101	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2480	0.2190	0.5508	∑ -Log(ρ[j])/n
RMSE	0.2659	0.2469	0.4259	√ ∑(y[j]-ρ[j])^{2}/n
Mean Abs Dev	0.1895	0.1594	0.2946	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0567	0.0638	0.2553	∑ (ρ[j]≠ρMax)/n
N	141	47	47	n

Confusion Matrix

Training

Actual	Predicted Count	
CDR 0/1	0	1
0	81	0
1	8	52

Validation

Actual	Predicted Count	
CDR 0/1	0	1
0	26	1
1	2	18

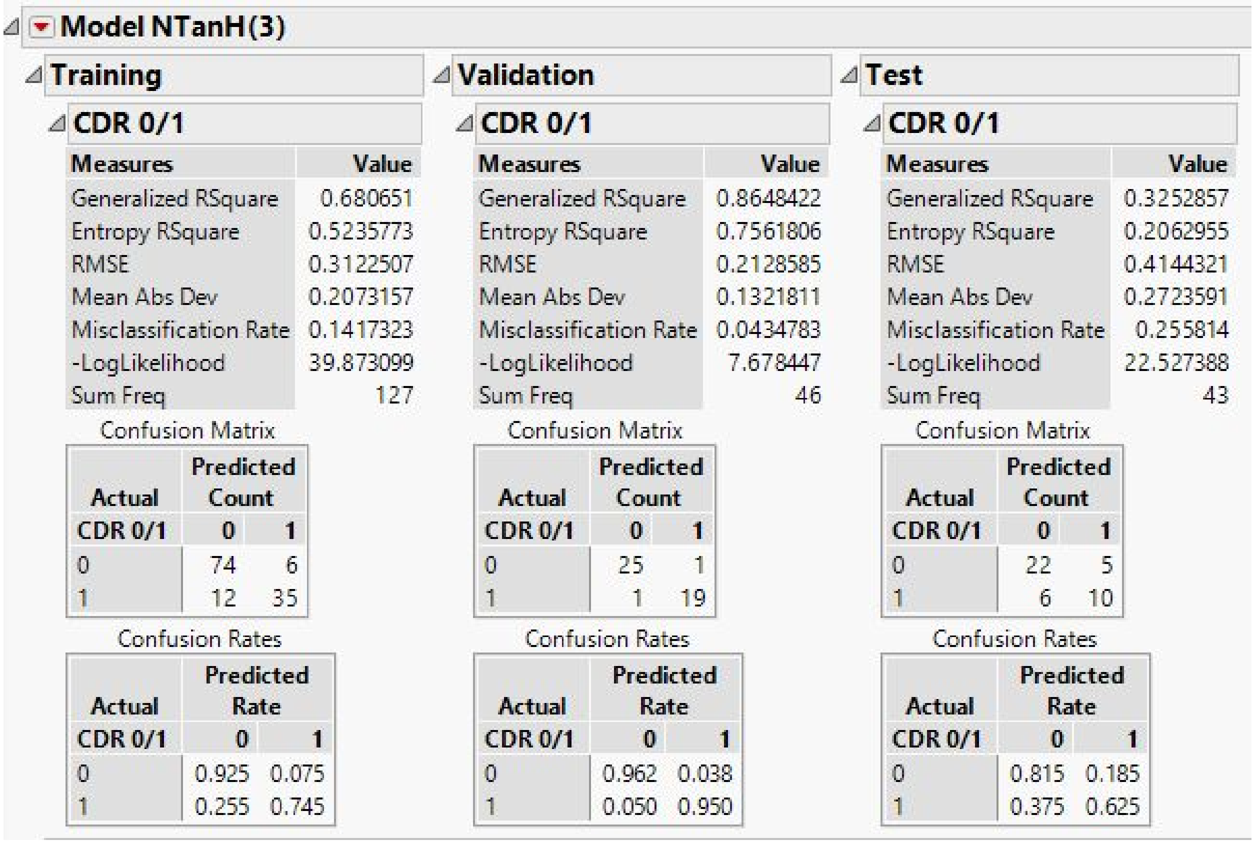
Test

Actual	Predicted Count	
CDR 0/1	0	1
0	24	3
1	9	11

Neural Network- Prediction

Again, only one level was considered for the neural network model due to the lack of complexity of the data set. All variables were considered in the Neural Network model. For the validation set, the misclassification rate was noted to be 75.6%. The confusion matrix shows a very small chance of false negatives & positives. Full details of the model results can be seen in figure 26 below.

Figure 26: Neural Network Model & Confusion Matrix Results for CDR- Prediction



Model NTanH(3)

Training

CDR 0/1

Measures	Value
Generalized RSquare	0.680651
Entropy RSquare	0.5235773
RMSE	0.3122507
Mean Abs Dev	0.2073157
Misclassification Rate	0.1417323
-LogLikelihood	39.873099
Sum Freq	127

Confusion Matrix

Actual	Predicted Count	
CDR 0/1	0	1
0	74	6
1	12	35

Confusion Rates

Actual	Predicted Rate	
CDR 0/1	0	1
0	0.925	0.075
1	0.255	0.745

Validation

CDR 0/1

Measures	Value
Generalized RSquare	0.8648422
Entropy RSquare	0.7561806
RMSE	0.2128585
Mean Abs Dev	0.1321811
Misclassification Rate	0.0434783
-LogLikelihood	7.678447
Sum Freq	46

Confusion Matrix

Actual	Predicted Count	
CDR 0/1	0	1
0	25	1
1	1	19

Confusion Rates

Actual	Predicted Rate	
CDR 0/1	0	1
0	0.962	0.038
1	0.050	0.950

Test

CDR 0/1

Measures	Value
Generalized RSquare	0.3252857
Entropy RSquare	0.2062955
RMSE	0.4144321
Mean Abs Dev	0.2723591
Misclassification Rate	0.255814
-LogLikelihood	22.527388
Sum Freq	43

Confusion Matrix

Actual	Predicted Count	
CDR 0/1	0	1
0	22	5
1	6	10

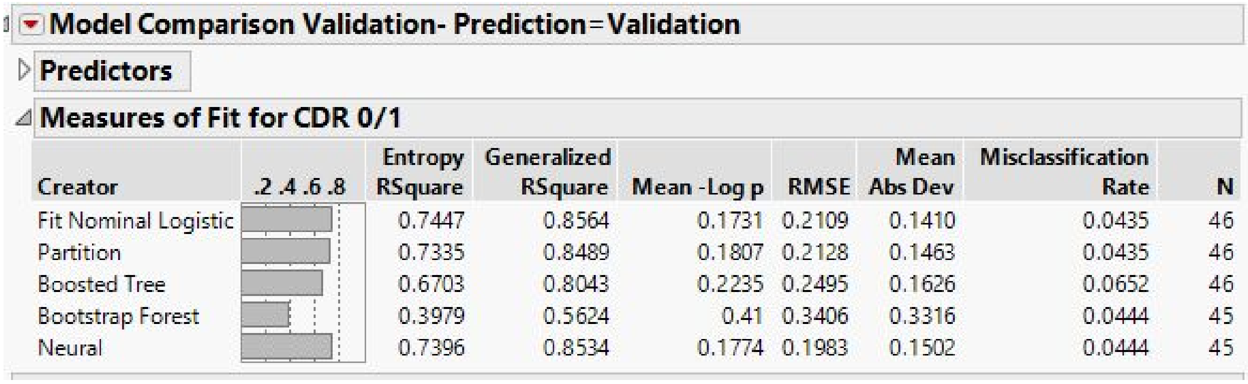
Confusion Rates

Actual	Predicted Rate	
CDR 0/1	0	1
0	0.815	0.185
1	0.375	0.625

Model Comparison & Selection- Prediction

Of the 5 models run for the prediction data set, the variable with the highest R-Squared value was the Fit Nominal Logistic Model, with R-Squared of 74.47%. This means that 74.47% of the variation seen in the model was explained by this model. This is fairly high, but there can still be some work done to improve the amount of variation explained in the model. Full results of the model comparison can be seen in Figure 27.

Figure 27: Model Comparison for 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

Prediction vs. Classification

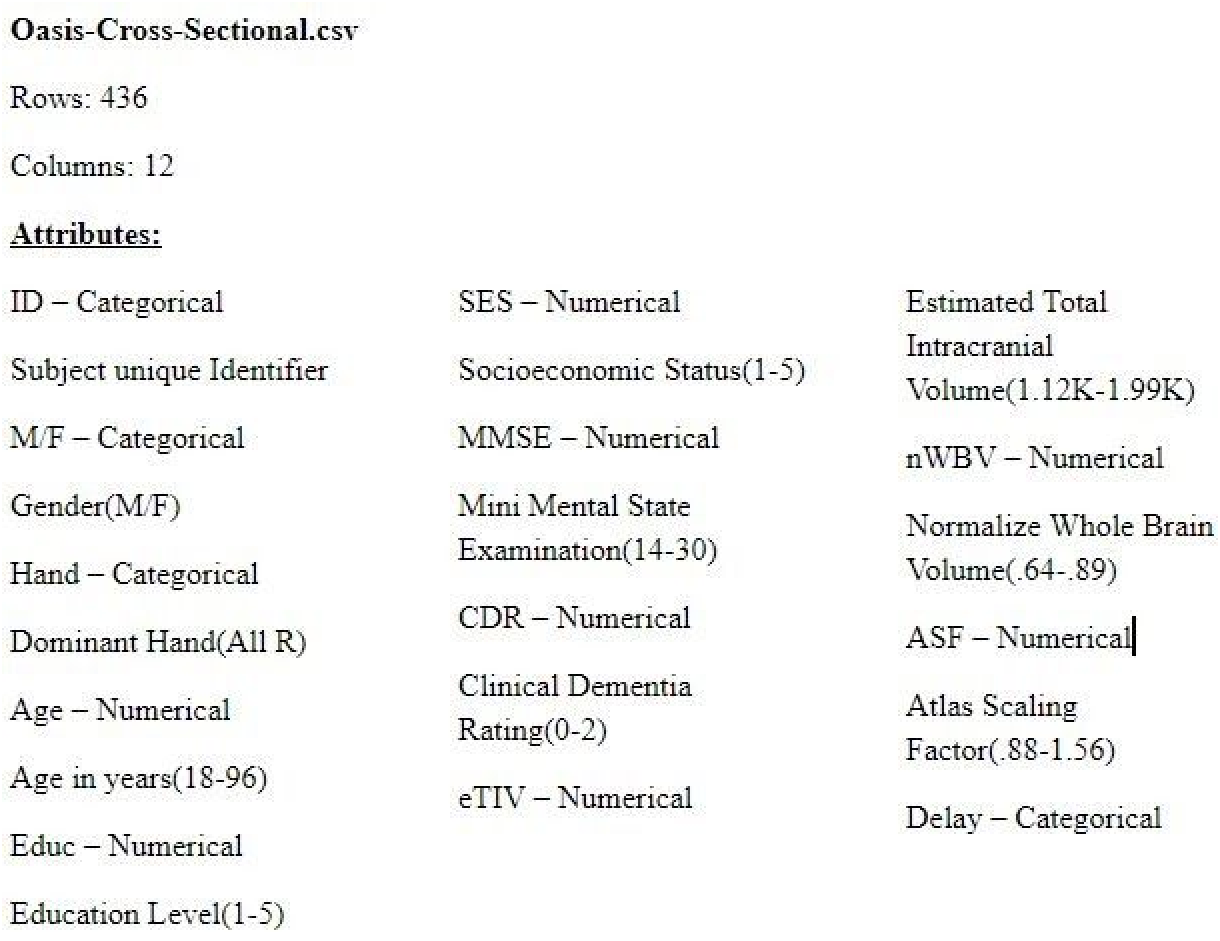
To compare this model to the classification model, the misclassification rates are considered. For all considered models in the prediction data set, all misclassification rates are noted to be much lower than those noted for the classification model, ranging from 4.35-6.52% (selected model, Fit Nominal Logistic, being 4.35% and meeting our desired <5% for this project). The lowest misclassification rate noted for the classification model was 20.93% on the Neural Network model. Overall, this indicates that the data set in its current state is not well suited to be a multi-level classification model and is better suited to be a two level 0/1 prediction model.

Conclusions & Recommendations

- **Note which variables remained significant vs. what visual analysis said**
    - **MMSE & nWBV were relevant**
    - **Others were not**
- **Based on analysis, data set is better suited for a prediction model than a classification model**
    - R-Squared of prediction model: 75.17%
    - Misclassification Rate of classification model: 20.93%
- **To improve classification abilities of the model:**
    - CDR sample should be leveled to include more patients from 0.5 to 2 categories (currently, 57% of data set represents CDR=0)
        - This will allow CDR=2 to be included in the model (a more serious form of Alzheimer's).
    - Excluded lines of data due to large amounts of missing prediction variables should be looked into to determine if those variables can be collected
    - If large amounts of predictor variables are still missing, impute data set and rerun model to see if classification rate improves

Appendix A

Summary of *MRI and Alzheimer's: Magnetic Resonance Imaging Comparisons of Demented and Nondemented Adults* by Jacob Boysen (taken from Kaggle.com)



Oasis-Cross-Sectional.csv

Rows: 436

Columns: 12

Attributes:

ID – Categorical

Subject unique Identifier

M/F – Categorical

Gender(M/F)

Hand – Categorical

Dominant Hand(All R)

Age – Numerical

Age in years(18-96)

Educ – Numerical

Education Level(1-5)

SES – Numerical

Socioeconomic Status(1-5)

MMSE – Numerical

Mini Mental State Examination(14-30)

CDR – Numerical

Clinical Dementia Rating(0-2)

eTIV – Numerical

Estimated Total Intracranial Volume(1.12K-1.99K)

nWBV – Numerical

Normalize Whole Brain Volume(.64-.89)

ASF – Numerical

Atlas Scaling Factor(.88-1.56)

Delay – Categorical