Building a Predictive Model for Diagnosing Dementia

Eric Gonzalez

Project Group 79 Email: egonzalez86@gatech.edu ID: 349

A final project for the class

Data Mining and Statistical Learning - ISYE 7406

Georgia Institute of Technology

November 28th, 2021

Building a Predictive Model for Diagnosing Dementia

Eric Gonzalez

Abstract

Dementia is a neurodegenerative disease that is increasingly prevalent in the U.S. and across the world. There is currently no cure and there is no one test for dementia. To address the gap in testing, this project aims to test a series of model with dementia data to identify which features and which model is best at predicting dementia in patients.

There were four models that obtained the best accuracy; however, the final model selected was logistic regression with stepwise variable selection using Bayesian information criterion (BIC). This model selected three variables, MR.Delay, SES, and CDR from the original set of features. This suggests that there were features that did not have relative importance or there was multicollinearity across the features. Dementia can impact anyone at anytime and this model can assist doctors and other healthcare professionals diagnose, treat, and eventually cure this devastating disease.

1 Introduction

Dementia is a neurodegenerative chronic illness that causes loss of memory, language, problem-solving and other neurological issues. It is defined as a deterioration of cognitive function beyond what may be expected. Common symptoms are memory loss, lapse in judgment, forgetfulness, limited social skills, and thinking abilities that inhibit day-to-day function. There are many types of dementia, with the most severe leading to Alzheimer's disease and eventually, death. While there are medications and therapies that can help manage symptoms, some causes are irreversible. Dementia does not have a cure.[1]

For those affected, dementia has a significant impact and social cost on families, friends, and caregivers. There are about 3 million cases per year with estimated care costs over $300+ billion per year. By 2050, the costs of dementia and Alzheimer's is expected to exceed $1 trillion per year.[2]

While nothing is directly linked with dementia, some common risk factors include:[3]

1. **Aging**: typically affects elderly individuals at a higher rate
    - Ages 65-69 (2 out of 100 people)
    - Ages 90+ (33 out of 100 people)
2. **Genes**: hereditary genes can increase the likelihood of dementia

- If a parent has a familial gene for dementia, there is a 1 in 2 chance a child develops dementia in their 50-60s

3. **Cognitive Reserve**: being able to cope with disease is lessened, which leads a higher likelihood for dementia for those with:
    - Low education-level
    - Low job complexity
    - Social isolation
4. **Demographic Factors**: social determinants of health play a role in dementia
    - Impacts women more than men
    - Affects African-Americans and Asians more than Caucasians
5. **Health Conditions**: certain comorbidities are linked to dementia
    - Cardiovascular issues
    - Hearing problems
    - Traumatic brain injury
    - Depression

There's no one test currently for dementia; however, people are diagnosed as early as 30-40 years old if there are visible signs of decline in thinking. A study performed in 2008 found that radiologists correctly diagnose and classify scans for Alzheimer's 65-95% of the time.[4] If caught early enough, dementia can be subdued and in some cases reversed.

2 Problem Statement

The goal of this project is to create the best predictive model for predicting dementia based on a set of features. The models created will address the following research questions about dementia:

1. Which model(s) performs best to predict dementia?
2. What variables are best at predicting dementia?
3. Is there any collinearity between the predicting variables from dementia? If so, do the models perform better with fewer features?

3 Data Sources

The dementia dataset is from Kaggle
(https://www.kaggle.com/shashwatwork/
dementiaprediction-dataset).

The dataset contains 918 participants with 14 predicting variables and 1 response variable (Group – Dementia status).

Models are built with a 80%-20% train-test split.

To prepare and clean the data, the following was performed:

1. Removed the ID features.
2. Removed the "Hand" feature as all were "R".

3. Changed “Converted” to “Demented” for the response variable “Group”.
4. Made the response variable and gender feature (M.F.) a binary (0 or 1).
5. Replaced NA variables for SES/MMSE with the mean.

4 Methods

The methodology for the project will be to test 11 models and compared the accuracy of each to determine which model performed best.
Models:

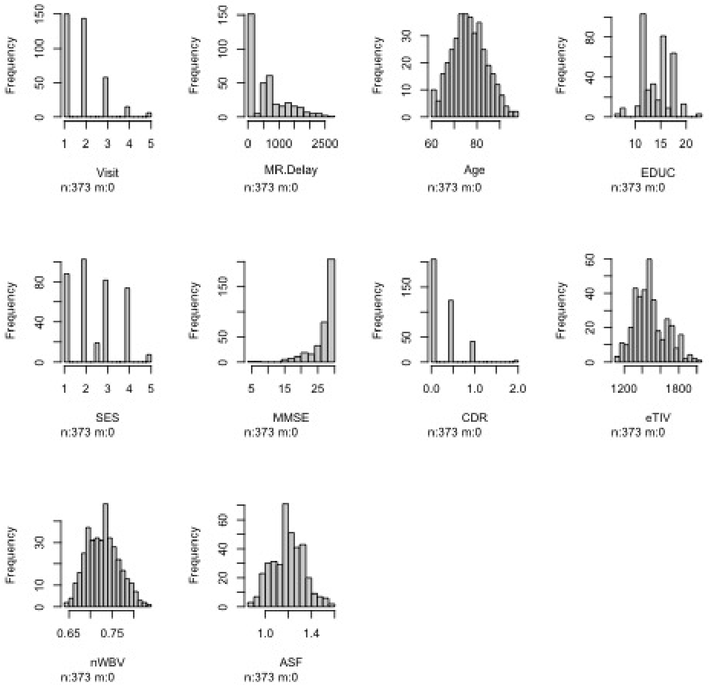
1. Logistic Regression
2. Logistic Regression Regression with Stepwise Variable Selection (AIC)
3. Logistic Regression Regression with Stepwise Variable Selection (BIC)
4. Naïve Bayes
5. Single Decision Tree
6. Random Forest
7. Generalized Boosting Model
8. QDA
9. LDA
10. KNN
11. PCA-KNN

Models will be built using the full set of data, then PCA and stepwise regression will be used for feature selection to potentially build a smaller model to see if a smaller model is more accurate in predicting dementia.

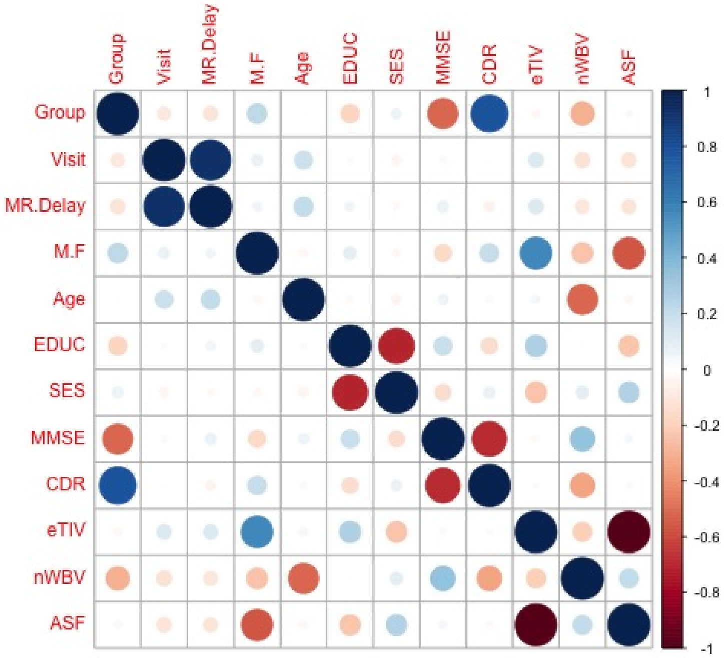
To test the accuracy of each of these models, the predicted binary response variable will be measured against the actual binary response variable with the total true positives + true negatives divided by the total number of observations.

5 Exploratory Analysis

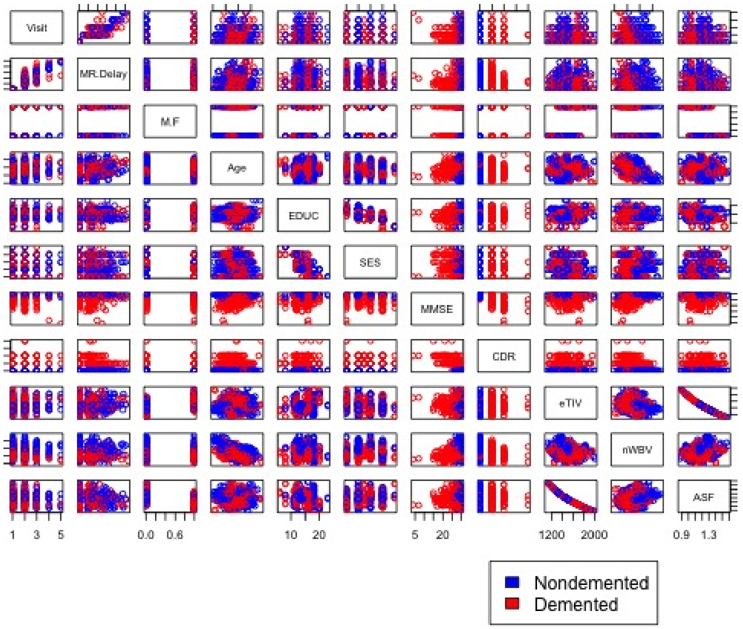
The data set is small overall with 918 rows and 14 columns, there are only 12,852 data points. This small sample size means the models built will have low bias and high variance. With such a small data set, this would normally may make it difficult for some of the models to accurately predict the response variable, Group; however, given we will be predicting a binary response, there may be higher accuracy for each of the models.



The histograms show that age, TIV, nWBV, and ASF have relatively normal distributions. Visit, MR.Delay, and CDR have right skewed distributions while MMSE has a left skewed distribution. Based on these distributions, there may need to be transformations performed on the variables if the model does not perform well. All this is to be expected given the size of the data set.



The correlation plot shows that MMSE is strongly negatively correlated, and CDR is strongly positively correlated with the response variable (Group). It also highlights that MR.Delay/Visit and eTIV/M.F have strong positive correlations with one another. Conversely, EDUC/SES and eTIV/ASF have strong negative correlations with one another.



The pairs plot shown highlights that MMSE and CDR show a good split between the demented and non-demented patients when combined with other predicting variables. MR.Delay and Age combined with MMSE, CDR, and nMBV also shows a good split of the response variable. The remaining variables may also be important and impactful for the model; however, it is unclear with the pairs plot.

6 Results and Findings

Results:

Model Results						
	Acc.	Sens.	Spec.	Pos. Pred. Value	Neg. Pred. Value	Log-Odds Ratio
Log. Reg.	95%	94%	**95%**	**94%**	95%	5.717
Log. Reg. w/ Step (AIC)	96%	97%	**95%**	**94%**	98%	6.436
Log. Reg. w/ Step (BIC)	**97%**	**100%**	**95%**	**94%**	**100%**	**Inf**
Naive Bayes	**97%**	**100%**	**95%**	**94%**	**100%**	**Inf**
Single Tree	**97%**	**100%**	**95%**	**94%**	**100%**	**Inf**
Random Forest	**97%**	**100%**	**95%**	**94%**	**100%**	**Inf**
Boosting	96%	97%	**95%**	**94%**	98%	6.436
QDA	96%	97%	**95%**	**94%**	98%	6.436
LDA	**97%**	**100%**	**95%**	**94%**	**100%**	**Inf**
KNN	60%	56%	62%	53%	65%	0.737
PCA-KNN	93%	97%	91%	88%	98%	5.678

Findings:

Of the 11 models, Logistic regression with stepwise regression (BIC), Naïve Bayes, Single Tree, Random Forest, and LDA all performed the best in accuracy as well as the other metrics. Given the poor performance by KNN and increased performance with PCA-KNN, higher accuracy of logistic regression with stepwise regression, random forest, and other models, it is clear that there was collinearity across the variables.

Of the four models that performed the best, random forest tends to work better than a single tree to reduce overfitting and with datasets where the predicting variables are categorical; however, these variables were not. While Naive Bayes performs better on smaller models, it assumes that all features are independent. Judging by the correlation map and pairs plot in the exploratory analysis, it is clear that this is not the case. Given this, **logistic regression with stepwise regression (BIC)** should be selected as the model for predicting dementia.

While radiologists and doctors do a decent job at diagnosing dementia, the use of a predictive model would help more accurately diagnose dementia (97% compared to 65-95%) and potentially diagnose earlier for treatment. If the logistic regression model had additional data on the health of the patient, such as medical claims or history, or had a larger sample size, the impact could be greater so that it could be more widely applied.

7 Appendix

Feature Selection, Cross-Validation, and Hyperparameter Tuning:

1. Logistic Regression with Stepwise Variable Selection (AIC):

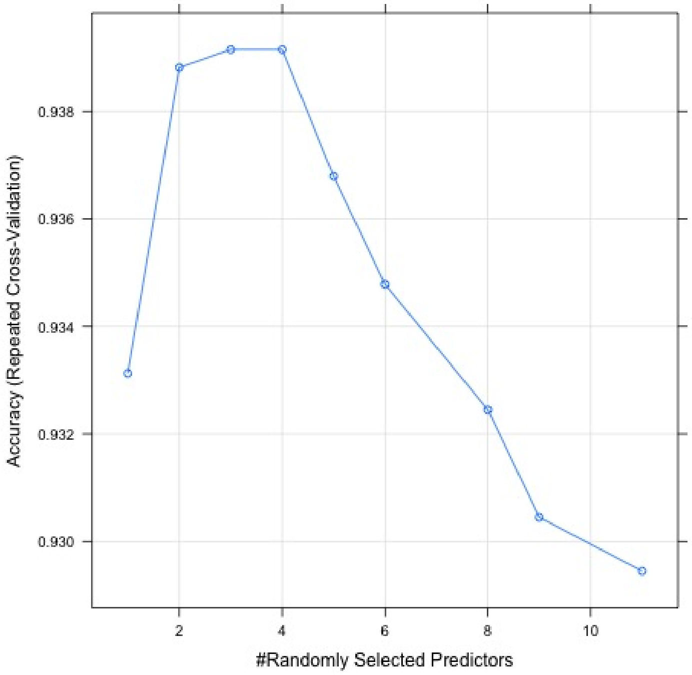
For this model, the Akaike Information Criterion (AIC) was used to select the best features from the original model. The features selected were: MR.Delay + M.F + SES + MMSE + CDR + eTIV + ASF

2. Logistic Regression with Stepwise Variable Selection (BIC):

   For this model, the Bayesian Information Criterion (BIC) was used to select the best features from the original model. The features selected were: MR.Delay + SES + CDR

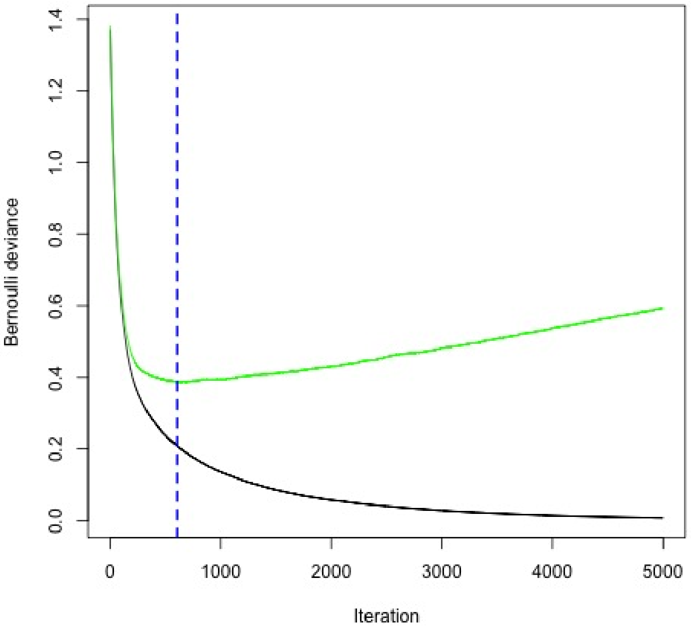
3. Random Forest:

   For this model, cross-validation was performed with accuracy as the metric to tune for. Three predictors were selected as it had the highest accuracy as shown on the chart below.

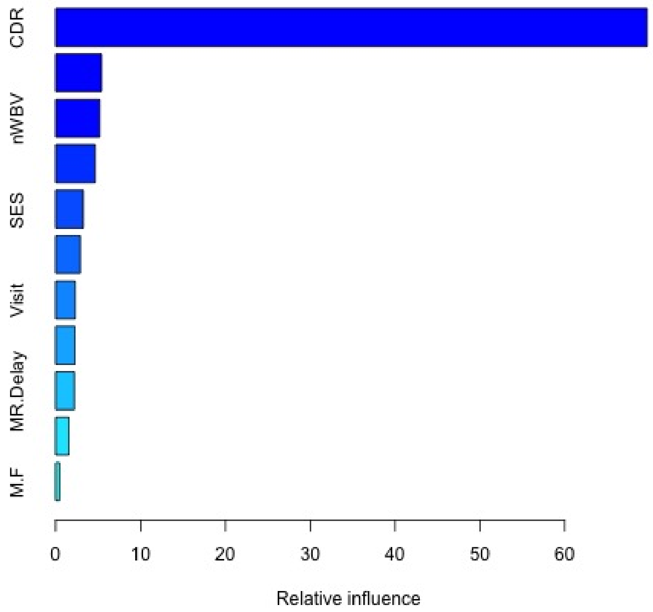


4. Generalized Boosting Model:

   For this model, the Bernoulli deviance was used for cross-validation to select the model with the best iteration. The best tree for prediction (604) is indicated by the blue dotted line below. The black line indicates the training Bernoulli deviance and the green line indicates the testing Bernoulli deviance. The tree selected minimizes the testing error on the cross-validation folds before the testing error increases over subsequent iterations.

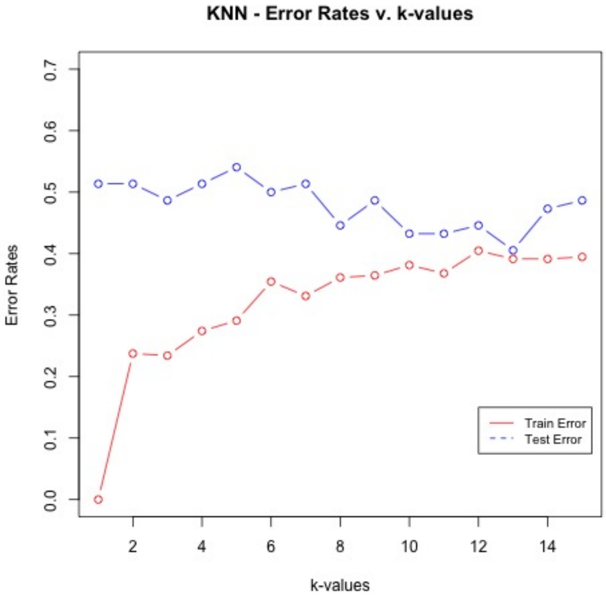

Bernoulli deviance
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
1000
2000
3000
4000
5000
Iteration

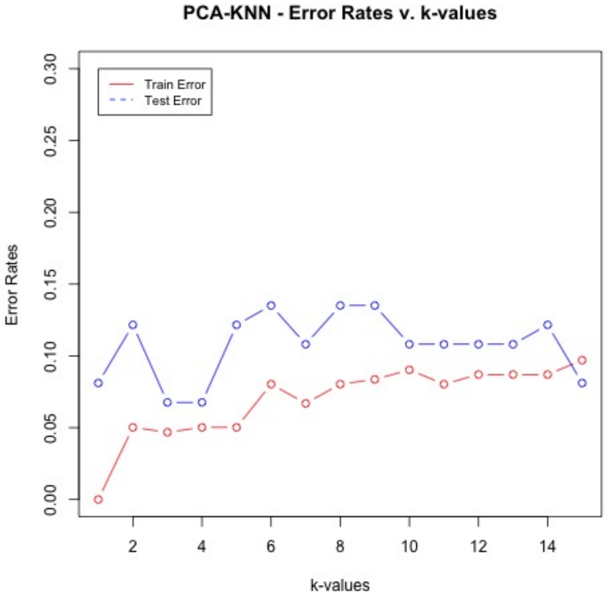
The variable with the largest influence is CDR and the variable with the least relative influence in M.F. This is surprising that M.F had the lowest relative influence as more women are affected than men by dementia.



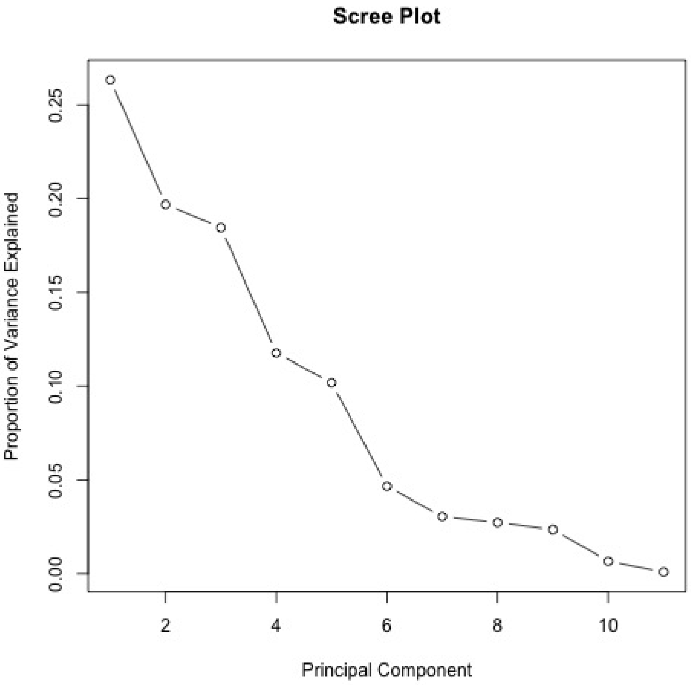
5. KNN & PCA-KNN

For these models, the k-value with best testing error rate was selected as indicated by the charts below.





For PCA, features were selected based on the magnitude of the coefficients. The scree plot below shows the eigenvalues on the y-axis and the number of factors on the x-axis. The ”elbow” of the curve or where the slope is leveling off indicates the number of factors to be used for the model, which is 7.



8 Lessons Learned

I enjoyed this project as it gives me hope and promise about the advancement of society, technology, and healthcare with the use of data science, AI, and predictive modeling. It also helped me understand that a lot of data, doesn't necessarily mean quality results as the best performing model reduced the features to 3 variables from the original set.

I took this class because I enjoyed my regression analysis course (ISYE 6414) and this was a continuation of those topics. Following this course, I feel that I have:

- Gained proficiency at coding in R
- Deepened my understanding of data modeling
- Learned new techniques for data mining
- Become a better data science practitioner

Thank you to Professor Mei and to all the TAs that made this course enjoyable!

References

[1] World Health Organization. Retrieved November 21st, 2021, from https://www.alz.org/alzheimers-dementia/facts-figures

[2] Alzheimer's Association. Retrieved November 4th, 2021, from https://www.alz.org/alzheimers-dementia/facts-figures

[3] Alzheimer's Society. Retrieved November 21st, 2021, from https://www.alzheimers.org.uk/sites/default/files/pdf/factsheet_risk_factors_for_dementia.pdf

[4] Klöppel S, Stonnington CM, Barnes J, Chen F, Chu C, Good CD, et al. . Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. Brain. (2008) 131:2969–74. 10.1093/brain/awn239