BIOST 546 Final Project

Katie McFarlane

3/10/23

Abstract

Current methods to diagnosis Alzheimer's disease are lacking, with many being invasive or expensive. Machine learning methods were applied to 360 cerebral cortex thickness measurements of 800 patients in an attempt to discern healthy patients from those with Alzheimer's disease in a more efficient and convenient way. Multiple methods were applied in hopes of finding the most accurate machine learning model; k-nearest neighbors, lasso regression, and boosted trees were all considered as potential techniques to address this problem. All three models performed with over 90% accuracy on the test set, which was comprised of 400 of the 800 patients. K-nearest neighbors performed the worst of the three methods, with a 91% accuracy when applied to the test set. Lasso regression had the best performance with 94.5% accuracy, and boosted trees returned 92% accuracy. While none of these methods are perfect, they provide a promising start to the process of redefining how we diagnosis Alzheimer's disease in individuals. These methods can be used by future researchers to learn more about the link between the cerebral cortex and Alzheimer's diagnosis.

Introduction

Problem of Interest

We were interested in applying machine learning methods to discern healthy patients from patients with Alzheimer's disease using cerebral cortex thickness measurements. In the past, the only definite method for diagnosis Alzheimer's disease was through conducting a post-mortem autopsy. While one can be diagnosed with 'probable' Alzheimer's disease, a lack of dependable methods of diagnosis can lead to misdiagnosis and cause stress in individuals experiencing Alzheimer's symptoms. Now, research has identified blood tests and various other scans that indicate signs of Alzheimer's disease with higher accuracy. However, research is still being conducted on Alzheimer's, and there is a need for less invasive, expensive, and time-intensive methods of diagnosis.[2]

Dataset Description

To address our problem of interest, we used a dataset with cerebral cortex measurements from 360 brain regions. There were 800 participants in the study, with diagnosis status known by the statistician for 400 of these subjects. Each measurement was recorded using the same units (mm). Variables were named generically, and the exact brain region these measurements were taken from was not recorded in or discernible from the dataset.[1]

Statistical Methods

Three models were used in this study: k-nearest neighbors, cross-validated lasso regression, and boosted trees. Note that for all models, the outcomes for the test set were not seen by the statistical

investigator, and the accuracy of models was tested by a separate individual who had access to the outcomes but did not influence model choice in any way.

First Model: K-Nearest Neighbors

The first method implemented was k-nearest neighbors. This model was chosen due to characteristics of the data available in this datset - knowing that cerebral cortex thickness is linked to Alzheimer's, it was believed that the diagnosis status of others with similar cerebral cortex thickness measurements will likely match their "neighbors."[3]

For this technique, the 400 data points with known outcome values were broken into a training set and a validation set, both with 200 data points each. All predictors in the training set were used to fit 50 k-nearest neighbor models, with k from 1 to 50. The validation set was then used to determine the number of nearest neighbors that returned the highest accuracy (defined as the lowest misclassification rate) on the validation set. This model with the best choice of k was then used to predict the classes for the test set of 400 individuals.

The data was not normalized for this method. This is due to the fact that all of the data measures cerebral cortex thickness, and as such, all data is on the same scale (mm).

Second Model: Lasso Regression

The second model was fit using cross-validated lasso regression. This model was chosen due to its balance between performance and model interpretability. Because this model performs variable selection, we anticipate using the results from this model to inform future studies into machine learning techniques regarding Alzheimer's diagnosis.

We used the entire training set to fit our lasso regression model with 5 folds. This model also used AUC as the loss, as we wanted the highest level of classification accuracy. We found the λ with the smallest cross-validated loss and used the model with that λ to predict the classes for the test set.

Third Model: Boosted Trees

The final model utilized boosted trees. This model was chosen due to being a tree-based model, which seems like a natural fit to this dataset: similar to k-nearest neighbors, we expected individuals with similar measurements to have the same diagnosis, and as such believed that, despite being a simpler technique, tree-based methods could perform well. Boosted trees in particular were chosen due to their performance compared to similar tree-based methods.

We used the entire training set to fit our boosted tree model with 500 trees, a shrinkage parameter of 0.01, and 2 splits in each tree. We then used this model to predict the classes for the test set.

Results

First Model: K-Nearest Neighbors

50 k-nearest neighbors models were fit, as specified in the Statistical Methods section. We examined the validation set accuracy for each value of k from 1 to 50 to pick the best value for k. See figure 1, where the dashed line indicates the chosen k.

We find the highest accuracy on the validation set for both k=19 and k=20, at 93.5%. Because both k=19 and k=20 return the same accuracy for the test and training sets, k=20 was chosen arbitrarily between the two. When we fit the model with k=20 to the 400 individuals in the test set, we found an accuracy of 91%.

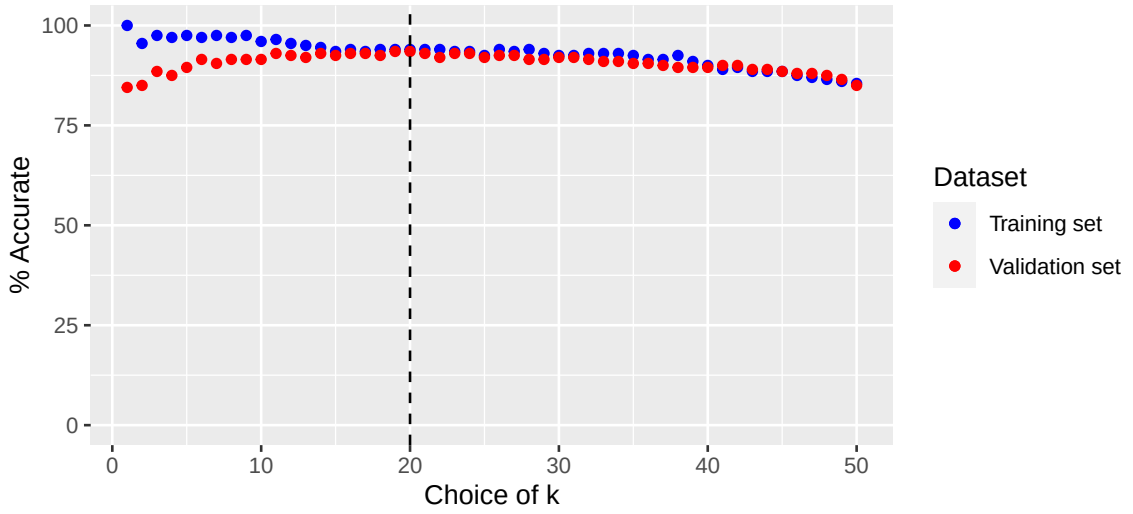

Figure 1: Training and validation set accuracy using k=1 to 50 nearest neighbors

Second Model: Lasso Regression

Next, we fit the cross-validated lasso model, and found that the log(λ) value that returns the highest accuracy on the training set is -3.16, as seen in figure 2: the left-most dashed line indicates the log(λ) chosen as the best. This model returned an accuracy of 93.75% on the training set.

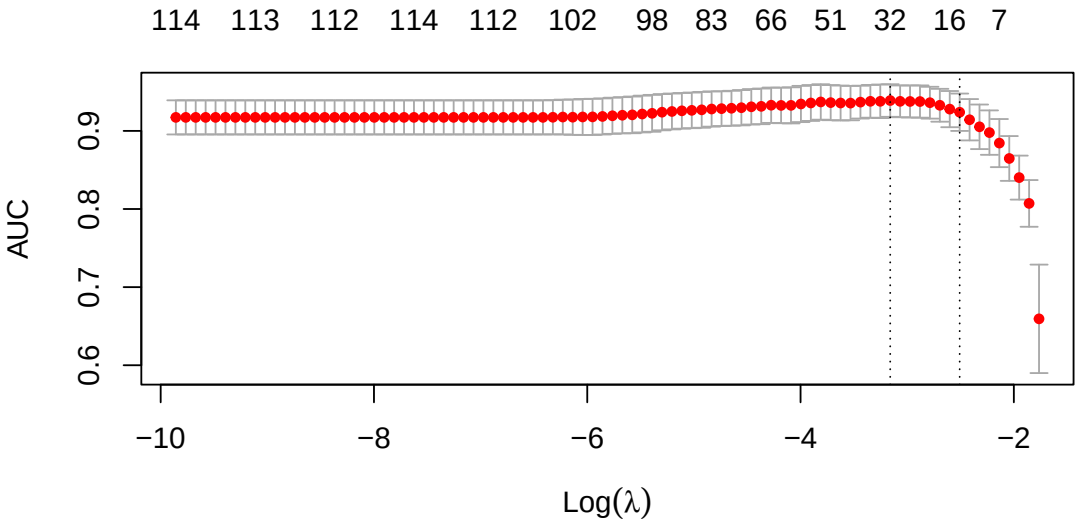

Figure 2: AUC of lasso regression models for log lambda values

We also found that using this best log(λ) value returned a model with 32 predictors, which allows for increased interpretability as compared to our k-nearest neighbors model, which included all 360 predictors.

When fit to the test set, we returned an accuracy of 94.5%. While we do not investigate it further, we recognize that the model using $\log(\lambda)$ one SE away from the chosen $\log(\lambda)$ would allow us to produce a model with only 13 variables with minimal change to the model accuracy.

Third Model: Boosted Trees

Lastly, we fit a boosted trees model and found that 143 of the 360 predictors have an influence in this model. All variables with a relative influence greater than 2 are plotted in figure 3 for illustrative purposes.

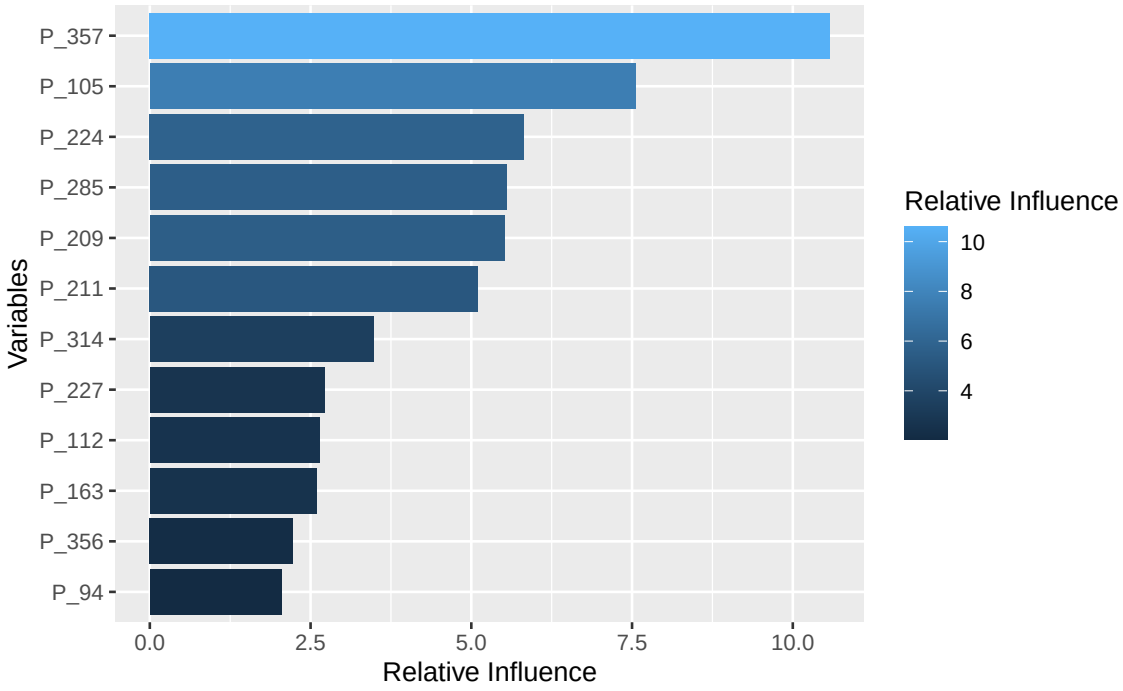

Figure 3: Variables with highest relative influence in the boosted tree model

We found that the boosted tree model has an accuracy of 97% on the training set. When fit to the test set, we returned an accuracy of 92%.

Model Comparisons

Every model has its pros and cons. The k-nearest neighbors model does not allow us to make inference on any estimated coefficients, but it performed fairly well (although the worst of these three models), and we can be fairly certain that we have chosen a useful number for k - although there was randomness in the choice of validation set, we do not expect better performance on the validation set if we had utilized k greater than 50, since we see in figure 1 that the accuracy peaks at k=19 and 20.

The lasso model returned the least number of parameters of these models, which allows us to make inference and gives some insight on what variables might be most useful in further studies. We also found a higher accuracy on the test set than we did on the training set, which was unexpected and may mean that the model is not overfitting the training data. This is preferred, since this means that this model is more likely to correctly classify new data that is not the training set. And, though we do not intend to further investigate this model, we note that the model using $\log(\lambda)$ one SE away from our chosen $\log(\lambda)$ produced a model with only 13 variables, which is less than half of the predictors chosen with the optimal $\log(\lambda)$ and would allow for even more ease of interpretation with this model.

The boosted model also performed variable selection, which helps us make inference, although nearly 4.5 times more variables remained in the boosted model when compared to the lasso regression model, which is less convenient for interpretation. However, we can use the information about relative influence to determine the top variables worth investigating further (as in figure 3). This model also performed slightly better than the k-nearest neighbors model.

While no model is perfect, for all three models, performance on the test set was above 90%. While lasso regression returned the highest accuracy, all three models performed well enough that further investigation into their uses for Alzheimer's diagnosis is justified.

Discussion

Limitations

There are a few limitations to this exploratory study. First, we used the same test data multiple times to get the test error for each model for comparison purposes, rather than one time as is preferred. There are also a multitude of other models that may work better to address this problem; only three were chosen in total to present in this report. Also, for the models that are reported in this project, we could have made further adjustments to certain model parameters (for example, by changing the type measure or the number of folds for the lasso regression, or the number of trees, interaction depth, or shrinkage effect for the boosted tree model) and checked against a validation set to find the optimal values for those parameters. Those changes may have resulted in more accurate classification. Finally, we also reiterate that this study did not take into account how physically far apart the measurements were taken within the cerebral cortex, which could be an important aspect of the relationship these variables have with the diagnosis of Alzheimer's disease.

Implications

Ultimately, we present three models that have high (>90%) accuracy of classifying Alzheimer's disease on our test set of 400 patients. We intend for these results to be used for both determining variables that are important to the relationship between cerebral cortex thickness values and Alzheimer's disease and for identifying the most accurate models in hopes of improving diagnosis in the future. We recognize that this study builds on the work of other research, and we hope that this work, too, can assist other works aiming to improve the lives and health outcomes of patients.[1] This is just one step towards improving the field for Alzheimer's patients and medical professionals worldwide.

References

1. Fischl, B. & Dale, A. M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proceedings of the National Academy of Sciences (2000). Available at: https://www.pnas.org/doi/10.1073/pnas.200033797.
2. How Is Alzheimer's Disease Diagnosed? National Institute on Aging (2022). Available at: https://www.nia.nih.gov/health/how-alzheimers-disease-diagnosed.
3. Lerch, J. P. et al. Focal Decline of Cortical Thickness in Alzheimer's Disease Identified by Computational Neuroanatomy. Cerebral Cortex (2005). Available at: https://academic.oup.com/cercor/article/15/7/995/388087.