Objectively Diagnosing Dementia

Rucha Acholkar, Norberto Mantohac, Charles Oh, Lois Yoo, Jessica Zhao
December 2023

Motivation

In this final project, our team aims to objectively explore the factors that are correlated with the onset of dementia. Our team harbors a curiosity in applying data to healthcare and are driven to identify the intricate relationships between lifestyle factors and health conditions in the context of dementia. The complexity of our project lies in the multifaceted aspects of dementia, where elements such as genetics, environmental factors, and individual health choices can correlate to the onset of this disease.

According to the National Institute of Aging, "dementia is the loss of cognitive functioning - thinking, remembering, and reasoning - to such an extent that it interferes with a person's daily life and activities" (NIH). Dementia is traditionally diagnosed subjectively by a mixture of cognitive and neurological tests, brain scans (CT, MRI, PET), psychiatric evaluation, genetic tests, and blood tests compared to our model's objective nature. Symptoms include memory loss, personality changes, inability to control emotions, difficulty speaking, reading, writing, understanding and expressing thoughts, etc. Additionally, about ⅓ of all people age 85 or older in the U.S. may have some form of dementia (NIH).

It is important to study the relationships between dementia and lifestyle/health factors, as strong insight into variable correlations of dementia can significantly contribute to preventative healthcare strategies. Early identification of potential risk factors for dementia allows for targeted interventions that can delay or mitigate the onset of dementia. Moreover, with these insights on dementia, individuals with susceptibilities to this condition - based on their lifestyle choices, genetic makeup, or health conditions - can receive personalized healthcare plans. By studying the intricate relationships between dementia and its potential contributing elements, we can support healthcare professionals in tailoring healthcare approaches to the specific needs of their patients. By studying the relationship between lifestyle, health, and dementia, we can also revolutionize preventative care and reduce the number of people susceptible to this condition.

Data

To explore this issue of dementia, we derived and utilized datasets from Kaggle. From the Alzheimer Features and MRI and Alzheimer's projects, we used the alzheimer.csv and oasis_cross-sectional.csv datasets respectively. These datasets were fit for our project because they contained both basic patient information as well as dementia-related statistics.

The alzheimer.csv dataset's basic patient information includes gender, age, SES (socioeconomic status), and education. It also contains features called MMSE (Mini Mental State Examination), CDR (Clinical Dementia Rating), eTIV (estimated Total Intracranial Volume), nWBV (normalized Whole Brain Volume), ASF (Atlas Scaling Factor), and Group, which classifies a patient as demented or nondemented. The oasis_cross-sectional.csv dataset has most of these same features, but it also specifies each patient's ID and dominant hand. MMSE refers to the score a patient got on the question-based assessment. CDR evaluates the severity of a patient's dementia through a numeric scale. eTIV and nWBV give information on brain size. Lastly, ASF is a one-parameter scaling factor that allows for comparison of the estimated total intracranial volume (eTIV) due to the various differences in human anatomy.

As for preprocessing, in order to create a well-prepped dataset for our project, we started off with fixing the EDUC column. We decided to rescale the EDUC column from the oasis_cross-sectional.csv to fit the alzheimer.csv's education column's 1 through 5 scale. Afterward, we merged the datasets through an outer join on all of the common columns between each dataframe. This means we also dropped unnecessary columns: Delay, MMSE, Group, ID, and Hand. With the remaining columns, we converted the Gender column's 'M' and 'F' values into binary values of 1 and 0. Then, to handle missing or null values, we decided to drop rows where the CDR column had null values as it is our independent variable. We also replaced any missing values in the SES column with the value of -1 due to its abundance. All of these operations gave us a dataset as such:

Table 1: Processed Dataset

	M/F	Age	Educ	SES	CDR	eTIV	nWBV	ASF
0	0	74	2.0	3.0	0.0	1344	0.743	1.306
1	0	55	4.0	1.0	0.0	1147	0.810	1.531
2	0	73	4.0	3.0	0.5	1454	0.708	1.207
8	1	74	5.0	2.0	0.0	1636	0.689	1.073
9	0	52	3.0	2.0	0.0	1321	0.827	1.329
11	0	81	5.0	2.0	0.0	1664	0.679	1.055
13	1	76	2.0	-1.0	0.5	1738	0.719	1.010
14	1	82	2.0	4.0	0.5	1477	0.739	1.188

The reason why we dropped MMSE is due to its strong negative correlation with our independent variable, CDR. They hold a correlation of -0.711793, so it could potentially lead to overshadowing other important predictors in our model. Additionally, by removing MMSE, we can focus on a broader range of variables that may offer unique insights into dementia progression, beyond what is captured by standard cognitive tests like MMSE, which can be affected by other patient factors such as hearing ability and English fluency. Lastly, we assured the validity of our merged dataset by calculating the overlap between them, which was under 0.27%. Since the overlap is very minimal, we can confidently use this dataset.

Additionally, for further exploratory data analysis, we plotted correlation graphs between all variables and discovered a noticeable relationship between eTIV and ASF. This relationship makes sense due to the fact that ASF scales eTIV. The correlation graph of nWBV and Age also reasonably reflects that humans' brain volume decreases as they age.

From an ethics standpoint, we can be confident that such concerns do not cloud this data as long as the patient information was collected with consent. We can also count on the fact that each patient connected to the information is displayed anonymously. However, one bias that we can acknowledge is the fact that our data does not consider non-dementia patients to compare with these dementia patients we analyze. If we could do so with more time and resources, our findings could be strengthened.

Analytics Models

We considered models used for binary classification because we wanted to predict whether a patient was diagnosed with dementia (1) or not (0) given a variety of objective features. Our reference for comparing model performance was a simple baseline model that predicted the most common outcome from our training set for all patients, which was that a patient did not have dementia.

The first model we constructed was a decision tree classifier with cross validation (CV CART). In general, CART with k-fold cross validation is a good way to perform binary classification because of its interpretability and ability to capture non-linear relationships, which is important if we hope our model can be used in applicable medical settings. The parameter we cross-validated was ccp_alpha, the improvement to the impurity function needed for a split to occur, and we used 5 random folds in the training set. The performance of this model in terms of accuracy, TPR, and FPR was much better than the baseline, and the tree corresponding to the model could be easily visualized. A drawback of this model, however, was that a singular CART model could have high variance if it were heavily dependent on our original dataset and random training-testing split. High variance in our context would mean that if we randomly removed or added patients to our dataset, our resulting CART model could be significantly different – for the generalizability of our conclusions, this is something we hoped to avoid.

To reduce variance and any potential issues caused by it, we built a random forest model and a gradient boosting model, both with cross validation. In the random forest model, we used cross validation on the max_features parameter to determine the optimal number of randomly-selected variables for the algorithm to examine at each split. This value was 5, out of the 7 independent variables in our data, and "hiding" information from the algorithm at each split allowed us to create a set of very diverse trees. Then, we combined by majority vote the results of 2000 CART models with our chosen parameters to get a final random forest model. Interestingly, the random forest model had a lower accuracy than the singular CART model, which suggests that the random forest model is more biased toward the training set data.

In the gradient boosting model, we used cross validation on the n_estimators and max_leaf_nodes parameters, and the results indicated that a large n_estimator value such as 2000 and a max_leaf_node value such as 10 were optimal. This is expected since the more boosting stages performed usually result in better performance, but it is not to say that the model is not prone to overfitting. Aside from our CART-based models, we also created a logistic regression model with a 0.7 p-value threshold in order to ensure that a higher probability value would lean towards the individual having an actual higher likely chance of having dementia. However, unsurprisingly, this model performed the worst (aside from baseline) in terms of accuracy and TPR due to the fact that logistic regression is better suited for discrete values, but our dataset is more complex with non-linearly related variables.

Table 2: Summary of Various Models' Performance Metrics

	Accuracy	TPR	FPR	PRE
Baseline	0.549180	0.000000	0.000000	0.549180
Decision Tree Classifier w/ CV (CART)	0.770492	0.654545	0.134328	0.800000
Random Forest w/ CV	0.754098	0.600000	0.119403	0.804878
Gradient Boosting w/ CV	0.786885	0.709091	0.149254	0.795918
Logistic Regression	0.631148	0.218182	0.029851	0.857143

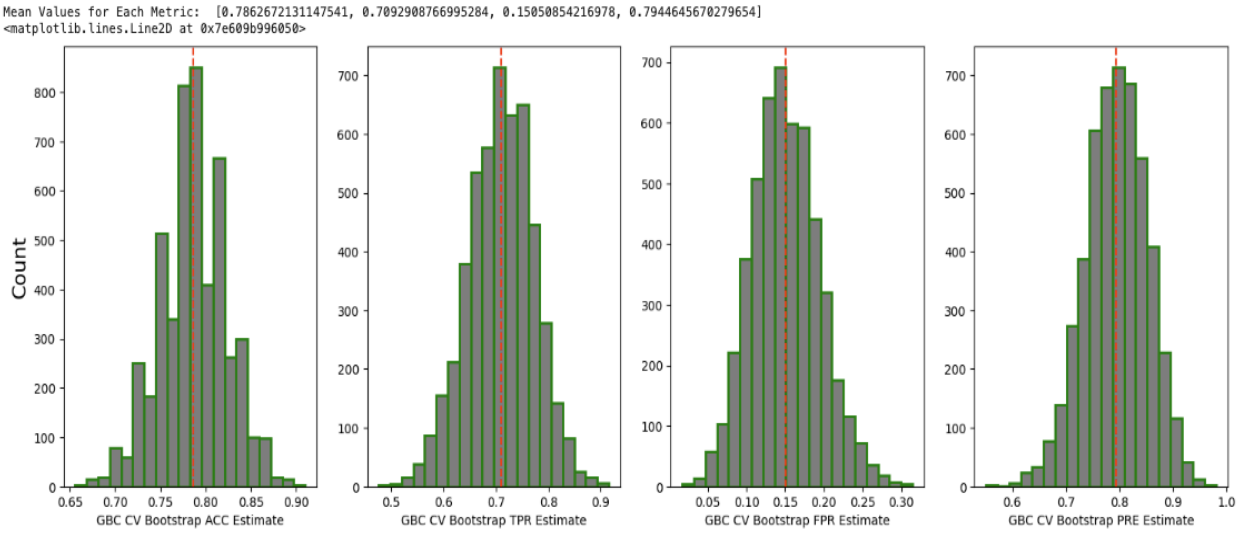
The results of the performance metrics for each of our model is shown above, and notable findings include: gradient boosting with cross validation model had the highest accuracy of

78.7% and the highest true positive rate (TPR) of 0.709, whereas the logistic regression model had the lowest false positive rate (FPR) of 0.0299.

When working with real world data, especially with health/medical data, it can be unclear which model is the best as the accuracies/TPR/FPR don't always align. In other words, the model with the highest accuracy or the highest TPR may not have the lowest FPR, such as in our analysis. In the context of this project TPR and FPR hold significant value in addition to accuracy. When analyzing model performance, it is ideal to maximize the TPR and minimize the FPR as much as possible. However, with real world data, the model performance is not always ideal and there sometimes needs to be a compromise between the trade-off of balancing TPR and FPR. In this project, a lower TPR means that people with dementia are getting classified as having no dementia, which can be dangerous as those individuals won't be able to receive the care they need. On the other hand, a high FPR means that more people are getting classified as having dementia, even if they actually don't have the condition, which can be dangerous for patients' mental health. Given that dementia is a serious condition, it is important that individuals get the necessary care. Again, in the context of this project, an ideal scenario would be where we have a high TPR and a low FPR; however, striking such an ideal balance with real world data is rare.

With this in mind, we determined our best model to be the gradient boosting with CV model because its accuracy and TPR were both the highest amongst all the models despite its FPR also being the highest, but it was still within a fairly reasonable range. With a high TPR, this indicates a low false negative (FN) trade-off, which we would want to minimize as it would mean that our model would be less likely to misdiagnose a patient having no dementia when they actually do.

Graph: Variability of Bootstrapped Performance Metrics



In order to assess how confident we are in our results, we performed 5000 bootstrapped samples using the gradient boosting w/ CV model in what is essentially bagging, as from the graphs above you can see the mean values and the variability of each performance metrics. If we were to make a 95% CI, we would confidently say that we would contain the true mean value of these performance metrics in our intervals 95% of the time, and notably, our current results of the gradient boosting model are already quite similar to the mean values above. Overall, gradient

boosting is a very flexible model, so there are still many ways to improve the performance by tuning more hyperparameter options such as tree depth and regularization methods to reduce overfitting, which can highly impact the TPR/FPR in accurately diagnosing a patient's dementia condition.

As a final addition to our set of models, we also experimented with generating a simple CART model that could be used as a basic pre-screen for dementia. We wanted to see if we could balance interpretability and model performance because although the models we created before were very accurate and good in terms of TPR and FPR, their logic and prediction processes could be difficult for patients and medical practitioners to understand, which should be avoided in any medical diagnosis. We tried implementing this in a few ways. First, we kept the same min_samples_leaf = 5 and ccp_alpha = 0.002 parameters as in the original CV CART model, and enforced max_depth = 3, which meant that we only wanted to make three or fewer comparisons before reaching a prediction. This model had eleven nodes and an accuracy of 0.639. Then, we built two more models without the max_depth parameter, one to increase min_samples_leaf and one to increase ccp_alpha, the motivation behind both being to create a tree with approximately ten nodes. Changing min_samples_leaf to 50 gave us a tree with fifteen nodes and an accuracy of 0.623; changing ccp_alpha to 0.01 gave us a tree with nine nodes and an accuracy of 0.623. From these tests and as expected, it was clear that a simple CART model would not perform as well as the complex one we had created originally. However, the performance of our simple CART models was similar to that of our logistic regression model, and we believe that the small sacrifice in accuracy was worth the gain we received in interpretability.

Impact

Through data cleaning/processing and modeling, we can explore the relationships and correlations between dementia and lifestyle or health factors. Our project, centered around building models to diagnose dementia, holds the potential to elevate the quality of healthcare. By integrating our model into clinical practices, healthcare professionals can leverage insights efficiently for early detection of dementia. When it comes to dementia, this early detection is crucial, as it can significantly delay the progression of symptoms, leading to better patient outcomes and quality of life. Moreover, the model can contribute to the development of personalized treatment plans, considering each patient's risk factors and health history.

A second impact of our project is that we can thoroughly understand the relationships between genetic makeup and lifestyle factors in addition to how they correlate with dementia. By investigating our findings, we can gain a stronger understanding of health choices that can prevent the development of dementia, educating the public about the disease. This information is especially important for those with a familial history of dementia, as this knowledge can help doctors and patients plan informed lifestyle decisions.

Furthermore, our findings from this project not only act as knowledge for the general public, but it can also support the broader healthcare system in gathering more insights into dementia. Our findings can inform policymakers in the public health sector. By identifying key factors contributing to dementia, researchers can dive deeper into understanding the backend mechanisms behind these correlations. This knowledge can lead to better-designed future research, which can lead to the development of new therapeutic treatment plans globally. Additionally, policymakers can use the findings to make public health campaigns to lower the prevalence of dementia in society.

Lastly, the societal impact of our project extends to the financial and economic domain.

The early detection of dementia can generate cost savings for healthcare systems globally. By lowering the need for extensive medical care, care facilities, and caregiver support for the treatment of dementia, our model can decrease the financial burden on families, decreasing public health budgets on a macro scale. Additionally, with a stronger understanding of lifestyle choices that can potentially lead to dementia, individuals can maintain their economic contribution in the workforce for a longer period of time.

Appendix

NIH citation source:
https://www.nia.nih.gov/health/alzheimers-and-dementia/what-dementia-symptoms-types-and-diagnosis

Link to code repository:
https://colab.research.google.com/drive/1PA4PRe7lO3QNrrvCPtyDmMsPjs9p4iJ8?usp=sharing

Links to Kaggle datasets:
https://www.kaggle.com/datasets/brsdincer/alzheimer-features
https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers