Predicting Alzheimer’s Diagnosis with Machine Learning Models

Erin Kwon, Charlotte Li, Larry Bermudez, Joshua Qian
Group 9

Introduction & Problem Description

Alzheimer's is a type of dementia that affects memory, thinking, and behavior. This is when brain cell connections and the cell itself deteriorates and dies, this then creates a plaque called beta-amyloid that forms in the brain tissue between nerve cells and causes the brain to shrink. Over the course of time, memories begin to fade and motor and mental functions decline. Main symptoms include memory loss and confusion. About 6.9 million people in the United States age 65 and older live with Alzheimer's disease. Among them, more than 70% are aged 75 and older. Of the more than 55 million people in the world with dementia, 60% to 70% are estimated to have Alzheimer's disease. As of right now, there is no cure for Alzheimer’s disease. However, there are medications available to temporarily alleviate and manage symptoms as well as slow down the disease progression.

It is quite challenging to diagnose Alzheimer’s as there is no definitive test and instead many assessments are taken to conclude a diagnosis. These assessments include: evaluating patient medical history, performing a mental status evaluation, conducting a physical exam, performing a neurological exam, laboratory testing, and further exams or tests. In this project we used an Alzheimer’s Disease Dataset from Kaggle that included testing and scan data that have been conducted on confirmed patients with Alzheimer’s as well as patients without Alzheimer’s. With this dataset we created a program that recognizes these patterns to help conclude diagnoses through the use of models such as LASSO Logistic Regression, Logistic Regression, and Neural Networks which explores the use of machine learning for the diagnosis of Alzheimer’s which is difficult to confirm.

Data Description

The Alzheimer’s Disease data sheet provides important information such as: demographic details, lifestyle factors, medical history, as well as cognitive and functional assessments. We are given a large sample group of 2149 patients. These parameters within the data sheet had ranges and caps. For example, regarding the age, all of the patients were aged 60 to 90, so the data is more controlled. This type of range is also seen in the other features such as alcohol consumption, physical activity, and diet quality. Aside from these ranges, the data also indicates the patient’s symptoms, such as: confusion, disorientation, personality changes, difficulty completing tasks, and forgetfulness. Along with the symptoms, assessments are taken and are a large part of when diagnosing Alzheimer’s. These assessments are included in the dataset to indicate if the patients have specific problems and where they test on a cognitive scale. Overall, there are 35 features in the dataset that give information about the patient, and we used 33 to help write our program, excluding the features ‘PatientID’ and ‘DoctorInCharge’ as we believed these were redundant. Furthermore, we identified that key features such as ‘MemoryComplaints’, ‘BehavioralProblems’, ‘ADL’, and ‘FunctionalAssessment’ were highly associated with Alzheimer’s diagnosis based on initial exploratory analysis. The dataset exhibited a slight class imbalance, with 64.63% of patients classified as non-Alzheimer’s and 35.37% classified as Alzheimer’s.

Methodology

Our team’s objective was to use either the ‘Diagnosis’ or ‘MMSE’ feature in our dataset as target variables for our trained models to determine whether a patient was diagnosed with Alzheimer's disease. The ‘Diagnosis’ feature is binary, making it fairly straightforward (1 representing Alzheimer’s and 0 representing no Alzheimer’s). The ‘MMSE’ feature is a score that determines an individual’s cognitive ability, where lower scores indicate cognitive impairment and higher scores suggest better cognitive function. During our initial exploration of the dataset, we realized that the dataset's features were not highly correlated with one another and this raised concerns about our models’ potential performance. However, we were advised to work with all features as low correlation coefficients do not directly imply that the features are uninformative. As suggested during our midpoint check-in meeting, we experimented with logistic regression models.

We first created a LASSO regularized logistic regression model with ‘Diagnosis’ assigned as the target variable. Since the ‘Diagnosis’ feature is binary, we were faced with a classification problem. The LASSO model allowed us to perform feature selection by shrinking coefficients of less informative features, effectively identifying the most significant predictors. We also programmed a typical logistic regression model with ‘MMSE’ assigned as the target variable. However, there were errors when we attempted to directly use the MMSE data and it was recommended that we convert the MMSE scores into binary values for classification. After some research, we decided to classify MMSE scores below 25 as 1 and scores of 25 and higher as 0, where 1 indicates a likelihood of Alzheimer's and 0 indicates its absence. To introduce diversity into our models, we developed a neural network model using the MLPClassifier algorithm from scikit-learn. The neural network included hyperparameter tuning with parameters such as the number of hidden layers, activation functions, solvers, and maximum iterations to enhance performance. For each model, 20% of the data was used as testing data while the remaining 80% was used as the training data.

Results & Analysis

<u>LASSO Regularized Logistic Regression Model Targeting Diagnosis</u>

Our LASSO regularized logistic regression model used the data from the dataset to create a model that effectively shrunk some coefficients to zero and performed feature selection by identifying the most important variables within our dataset. From this model, we were able to come to the conclusion that our best C value is the first value in our initial chosen interval. Afterwards, we performed another round of modeling of lower values to see if we could achieve a better C value. From the second model, we were able to see that our original C value was the best one, being 0.8051192553810355. With the best C value, we created our best fit model. The most significant features identified by the model were 'MemoryComplaints', 'BehavioralProblems', 'ADL' (Activities of Daily Living Score; measured from 0 - 10 where lower scores indicate greater impairment), and 'FunctionalAssessment' (measured from 0 to 10 where lower scores indicate greater impairment), which strongly align with clinical patterns in Alzheimer's diagnosis. This insight highlighted the importance of these features as these were the main features that the model used to come to a conclusion.

Checking the accuracy of this model, we were able to get an accuracy rate of 84.7%. We created a confusion matrix to visually evaluate the performance of the predicted label and the true label. By analyzing the table, we were given the result that the model correctly identified 70.4% of patients with Alzheimer's (true positive rate) and 91.7% of patients without Alzheimer's (true negative rate). To further evaluate our model, the cross entropy loss was calculated to see the measure of how well the model's prediction matched the actual outcome.

The outcome was relatively low at 0.394, which indicated that the model was predicting probabilities close to the true probabilities since a cross entropy loss value less than 0.5 is ideal. We also created a calibration curve with the true probabilities to visually represent the relationship between the true probabilities and the model predictions. The calibration curve indicated that the model's prediction aligned relatively closely with the ideal calibration curve. The cross-validation score for this model is 80.5%, supporting the idea that the model performs generally well.

Logistic Regression Model Targeting MMSE Score

The model that appeared to have the best accuracy is our simple logistic regression model with 'MMSE' as our target variable, with an accuracy of 87.2%. However, after the conversion of our MMSE scores to binary values, we observed that the data was imbalanced, with significantly more scores below the 25 threshold than above it. This prompted us to look at the true positive and true negative rate of this model by analyzing its confusion matrix. The true positive rate was 99.7%, while the true negative rate was only 3.57%. As expected, the model excelled at predicting patients with MMSE scores below 25 but performed poorly when predicting patients with MMSE scores above 25. The mean cross-validation accuracy for our logistic regression model across 3 folds is 83.7% which is generally considered satisfactory for simple models such as this one. However, to address the imbalance in our data, we decided to add the class_weight='balanced' hyperparameter to the model.

This new model had an accuracy of 58.1%, which was much lower than our previous model. The new true positive rate dropped to 54.5% while the new true negative rate increased to 82.1%. Although the updated model performs better at identifying patients with MMSE scores greater than or equal to 25, it now struggles more with correctly identifying patients with scores below 25. The new mean cross-validation accuracy across 3 folds decreased to 61.4%, suggesting that this new model is not performing as well as expected. We wanted to see how each feature was accounted for in this new model and saw that 'ADL', 'MemoryComplaints', and 'BehavioralProblems' were the main features it identified. There was also a wider range of features it also identified such as 'CholesterolTotal' and 'Smoking' among others. While in the LASSO Logistic Regression model they were not key features, this model might provide further insight as to what other factors can lead to an increased chance of developing Alzheimers.

However, according to this Feature Importance graph, ‘FunctionalAssessment’ appears to be insignificant, despite being important in the previous LASSO Logistic Regression model. This may be due to differences in correlation between ‘MMSE’ and ‘FunctionalAssessment’ in comparison to ‘Diagnosis’ and ‘FunctionalAssessment’.

Neural Network Targeting Diagnosis

Our neural network model had an initial accuracy of 74.9%, lower than both our previous regression models. By observing its confusion matrix, the true positive rate was 67.6%, while the true negative rate was 78.5%. This implies that although the neural network has a lower accuracy than the simple logistic regression model, it has a better balance of correctly identifying patients with Alzheimer’s and those without Alzheimer’s in comparison to it. The loss curve drops significantly within the first 100 iterations, suggesting that the model is learning effectively through the data. The mean cross-validation accuracy for our neural network model across 5 folds was 73.5%, performing relatively well but there is still room for improvement.

After these results we decided to perform hyperparameter tuning using Grid Search with different parameters such as the hidden_layer_sizes (architecture), activation, solver (optimization method), and maximum number of iterations. Afterwards, the mean cross-validation accuracy increased to 79.2% and the general accuracy increased to 82.3%. The true positive rate increased to 69.7%, while the true negative rate increased to 88.5%. We then decided to change the number of folds from 5 to 3 to see how the performance of our model would differ. With the number of folds being set to 3, the mean cross-validation accuracy increased slightly to 79.3% and the general accuracy increased to 84.4%. The true positive rate also increased to 75.4%, while the true negative rate increased to 88.9%. This increase in accuracy as a result of decreasing the number of folds could be because of the larger training dataset assigned to each fold, encouraging the model to learn more effectively from the data. We were unable to improve the accuracy as much as we would have liked as the hyperparameter tuning took longer to run in our code than anticipated.

Model	Accuracy	True Positive Rate	True Negative Rate	Cross-Validation Accuracy
LASSO Regularized Logistic Regression	84.7%	70.4%	91.7%	80.5%
Logistic Regression	87.2%	99.7%	3.57%	83.7%
Logistic Regression Model (class_weight='balanced')	58.1%	54.5%	82.1%	61.4%
Initial Neural Network	74.9%	67.6%	78.5%	73.5%
Hyper-Parametrized Neural Network (K=5)	82.3%	69.7%	88.5%	79.2%
Hyper-Parametrized Neural Network (K=3)	84.4%	75.4%	88.9%	79.3%

Conclusion & Additional Thoughts

If neural networks are preferred, the hyper-parametrized neural network with K=3 performs the best in this category. If we had more time to run the code, there could also be better neural network models with different hyperparameters. If we are prioritizing general accuracy and cross-validation accuracy then the simple logistic regression model performs best. On the other hand, the LASSO regularized logistic regression model demonstrates a better balance between the true positive and true negative rate. Overall, we came to the conclusion that the LASSO regularized logistic regression model was the best model out of the three as it was the most well-rounded. It has high accuracy and cross-validation accuracy as well as having balanced true positive and negative rates.

For research in the future, we could explore whether additional features not present in the dataset could better contribute to the doctor's diagnosis. Features like brain scans such as MRI's could possibly help the program come to a better conclusion, as the more data present curates more of an accurate diagnosis. As the diagnosis is dependent on the doctor's decision, it may be subject to flaws due to the presence of human error and emotions. These two significant factors could subject the patient to bias where the doctor could possibly misdiagnose the patient. We could also analyze the features further to determine whether the diagnosis was consistent for similar or identical features. Analyzing diagnoses with similar correlated features would train the program to understand which features result in an Alzheimer's diagnosis in some individuals in comparison to others. It is an exciting time to explore the power of machine learning and its capabilities across a multitude of industries and the healthcare sector is certainly an area of high interest. With the growing use of machine learning and AI, its implications within medicine are actively revolutionizing patient care by enabling accurate diagnosis through predictive models.