Predictions on Diagnosis of Alzheimer’s Disease

Research Area

Alzheimer’s Disease (AD), the leading cause of dementia, is a progressive neurodegenerative disorder that affects millions globally. It is characterized by memory loss, cognitive decline, and functional impairment. The disease’s prevalence is expected to triple by 2050, making it a pressing global health challenge. In Canada alone, 9.7% of women and 6.9% of men aged 65 and older were affected by dementia as of 2022. These figures are projected to rise to 15.3% for women and 10.6% for men by 2050, with Alzheimer’s Disease accounting for 60% to 80% of dementia cases (Elflein, 2024). This rising prevalence underscores the need for effective early detection and intervention strategies. Early diagnosis of AD is critical as it enables timely interventions that can slow disease progression, improve patient outcomes, and reduce caregiver burdens (Alzheimer’s Disease International, 2022). Detecting AD at its early stages offers opportunities for therapeutic interventions, lifestyle modifications, and long-term care planning. However, diagnosing AD accurately remains a challenge due to its multifactorial nature. The disease is influenced by a complex interaction of genetic, lifestyle, and medical factors, requiring thorough analysis to pinpoint reliable diagnostic markers.

Machine learning (ML) techniques have emerged as powerful tools for addressing complex medical problems, including Alzheimer’s Disease diagnosis. ML enables the integration and analysis of large, multidimensional datasets, uncovering patterns that are often undetectable through traditional clinical methods. This project leverages an extensive dataset containing demographic, lifestyle, medical, and cognitive assessment information for 1,504 patients. The dataset includes features such as the Mini-Mental State Examination (MMSE), which is widely recognized as a reliable cognitive assessment tool (Folstein et al., 1975); Activities of Daily Living (ADL), which helps assess the functional decline characteristic of AD; and Behavioral Problems, which have been identified as significant predictors of AD progression (Barrett-Connor et al., 2008).

The goals of this project are twofold: (1) to develop a predictive model for Alzheimer’s disease diagnosis using ML techniques and (2) to explore the factors most strongly associated with AD. Predictive modeling can serve as a decision support tool for healthcare providers, enabling them to identify at-risk individuals earlier and more accurately. Additionally, analyzing the importance of features within the model provides insights into modifiable risk factors, such as physical activity and diet quality, which could inform preventive strategies (Stefaniak et al., 2022). By addressing these objectives, this study contributes to the growing body of knowledge on AD diagnosis and management. It highlights the potential of combining clinical expertise with ML to improve health outcomes and reduce the societal burden of AD. This research not only aligns with the global push for early diagnosis and treatment of dementia but also underscores the critical role of advanced analytics in tackling healthcare challenges.

Statistical Analysis

The dataset we are working with includes health information for 1,504 patients, each identified by a unique Patient ID. The primary outcome of interest is the diagnosis status of Alzheimer's Disease, represented as a binary variable (0: No, 1: Yes). This dataset provides a wide range of attributes, including demographic information (age, gender, ethnicity, and education level), lifestyle factors (BMI, smoking, physical activity, and diet quality), medical history (diabetes, cardiovascular disease, and family history of Alzheimer’s), clinical measurements (blood pressure and cholesterol levels), cognitive assessments (MMSE and functional assessment scores), and reported symptoms.

Data Cleaning

Before conducting statistical analysis, it is essential to ensure the dataset is clean and optimized for predictive modeling. In this project, we began by checking for missing values in the training dataset and confirmed that no NULL values were present. Next, we identified and removed two features deemed irrelevant to the predictive modeling task: PatientID and DoctorInCharge. These features were excluded because they are unique identifiers or administrative details that do not contribute to the prediction of Alzheimer’s diagnosis. We retained the remaining 33 variables that may have meaningful relationships with the target variable with 1,504 observations.

Exploratory Data Analysis (EDA)

Following data cleaning, we performed exploratory data analysis (EDA) to uncover insights into the dataset and its relationship with the target variable, Alzheimer's diagnosis. The steps involved numerical and categorical variable exploration, outlier detection, and visualization to understand feature relationships, distributions, and correlations.

We began with a summary of the training dataset to understand the overall structure, including ranges, means, and distributions of both numerical and categorical features. Subsequently, we looked at the relationships between the most relevant numerical variables and Diagnosis. We deemed age, a discrete numerical variable, to be highly relevant based on existing literature and visualized this relationship by grouping patients by age and calculating the average diagnosis value (indicating the proportion of Alzheimer-positive cases in each age group) (Guerreiro & Bras, 2015).

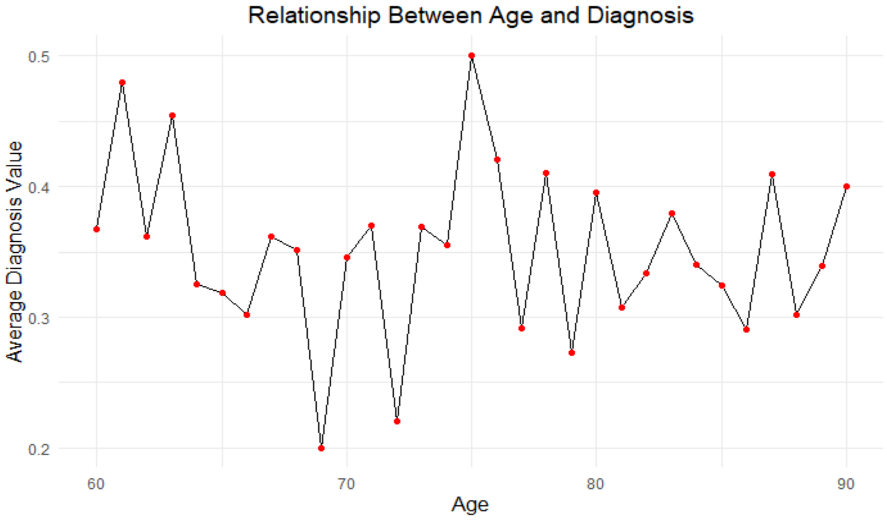


Figure 1: Visualizing Relationship between Age and Alzheimer's Diagnosis

A line plot was used to display this trend. As seen in Figure 1, there is a fluctuating but generally more consistent trend in the average diagnosis value as age progresses. This indicates that individuals in higher age groups have a more consistent likelihood of being diagnosed with Alzheimer's. The variability in the line plot may reflect natural population differences or sampling variation. Overall, from Figure 1, there is no significant trend to be drawn between an increase in age and Alzheimer's diagnosis.

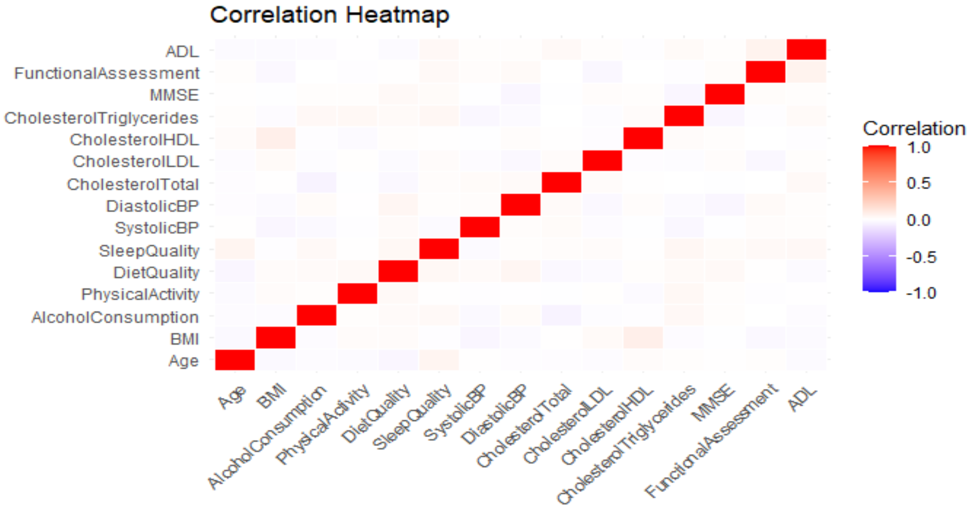


Figure 2: Heatmap of Correlation between Numerical Variables

Additionally, we examined the correlations between the numerical variables for potential multicollinearity. These variables include Age, BMI, Alcohol Consumption, and several health and cognitive metrics. A correlation matrix was computed to explore the relationships among these features, which could impact model performance and feature selection. The correlation matrix was visualized using a heatmap, with color gradients representing the strength and direction of correlations (from -1 to 1). Figure 2 demonstrates generally weak correlations among the numerical variables in the dataset, with no correlation coefficients exceeding 0.3. This is evident from the unsaturated colors, which indicate low correlation values.

The last analysis we performed on numerical variables is outliers detection. We identified outliers for numerical variables using the Interquartile Range (IQR) method. Each variable's lower and upper bounds were calculated, and the number of outliers was summarized in a table. Overall, there were no outliers for the numerical variables.

Next, we performed an analysis of categorical variables. Bar plots were created to examine the relationship between categorical variables and Alzheimer's diagnosis. Key features analyzed included: demographic factors (Ethnicity and Education Level), lifestyle and behavioral factors (Smoking, Family History of Alzheimer's, and Memory Complaints), and cognitive and symptomatic variables (Confusion, Personality Changes, and Forgetfulness). Each bar plot stratified the data by diagnosis, revealing clear proportional distinctions in variables like Behavioural Problems and Memory Complaints, which appeared significantly more common in Alzheimer-positive individuals.

Furthermore, we explored the relationship between Gender and Diagnosis specifically and changed the variable from text to numerical values (0 = Male, 1 = Female). The average diagnosis value for males and females was calculated and displayed using a bar chart in Figure 3.

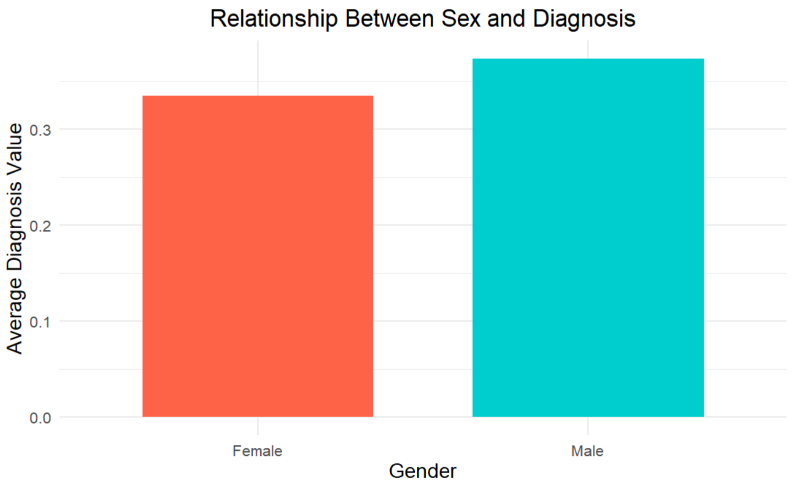


Figure 3: Diagnosis Proportion for Males and Females

As illustrated in Figure 3, males exhibit a slightly higher average diagnosis value compared to females. This indicates a marginally increased prevalence of Alzheimer's diagnosis among males in this dataset. However, the difference is not substantial and sex may not be a significant variable in estimating Alzheimer's diagnosis.

Overall, the exploratory data analysis provided valuable insights into the Alzheimer's dataset. For numerical variables, age was explored as a key factor, but no significant trend was observed linking increasing age to higher Alzheimer's diagnosis rates. A correlation analysis revealed weak relationships between numerical features, indicating minimal multicollinearity and suggesting the features contribute independently to the prediction. Outlier analysis confirmed the absence of extreme anomalies in the numerical variables.

For categorical variables, distinct patterns emerged. Variables such as Behavioral Problems, Memory Complaints, showed clear proportional distinctions between diagnosed and undiagnosed groups, underscoring their potential predictive value. Additionally, a slight difference was observed in diagnosis prevalence between males and females, with males showing marginally higher rates, although this difference was not substantial. Overall, the EDA confirmed that the dataset is clean and free of multicollinearity or significant outliers, setting a strong foundation for feature selection and modeling.

Feature selection

For feature selection, we employed a combination of statistical and machine-learning techniques. Firstly, we applied the Least Absolute Shrinkage and Selection Operator (LASSO) to identify a subset of variables most predictive of Alzheimer's diagnosis. LASSO is particularly effective in handling multicollinearity and improving model interpretability by shrinking less important variable coefficients to zero (Tibshirani, 1996). Using cross-validation with a binomial family model, we determined the optimal regularization parameter to balance model performance and complexity. Variables with coefficients above a defined threshold were selected, resulting in the following reduced set of predictors: MMSE (Mini-Mental State Examination), Functional Assessment, Memory Complaints, Behavioral Problems, and ADL (Activities of Daily Living).

The MMSE is widely recognized as a critical diagnostic tool for cognitive impairment and a strong predictor of Alzheimer's disease (Folstein et al., 1975). Similarly, ADL scores, which assess functional capabilities, are crucial for understanding the progression of Alzheimer's and its impact on daily living (Barrett-Connor et al., 2008). Features such as Behavioral Problems and Memory Complaints are consistently highlighted in the literature for their strong associations with Alzheimer's diagnosis (Stefaniak et al., 2022).

To refine the model further, we employed stepwise selection methods using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). These methods iteratively add or remove variables to identify the most parsimonious model. BIC, which imposes stricter penalties on model complexity compared to AIC, tends to favor simpler models. With AIC stepwise selection, one additional feature - Sleep Quality - was selected. However, its coefficient estimate was not statistically significant (p-value = 0.116). BIC stepwise selection reaffirmed the importance of the five predictors identified by LASSO, emphasizing their robustness.

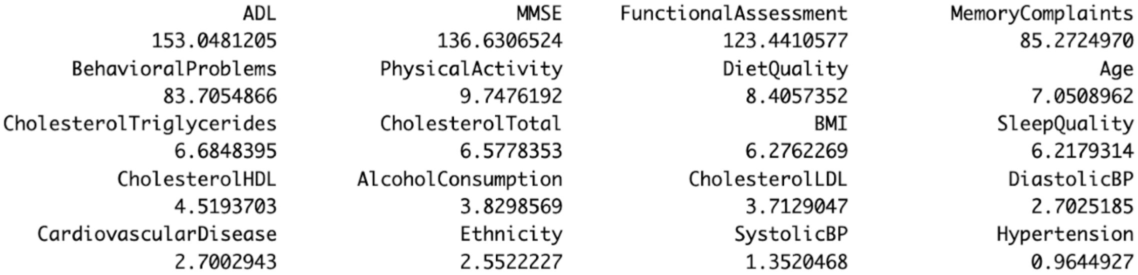
As an alternative perspective, we built a decision tree using all available predictors. Decision trees inherently handle variable selection and rank feature importance based on their contributions to tree splits. This analysis corroborated the relative importance of features identified by LASSO and stepwise selection, highlighting MMSE, ADL, and Functional Assessment as the most critical predictors.

By leveraging LASSO, stepwise selection with AIC/BIC, and decision tree analysis, we identified a consistent set of five core predictors: MMSE, Functional Assessment, Memory Complaints, Behavioral Problems, and ADL. These variables form the foundation for building predictive models in subsequent steps.

Machine Learning Algorithm

In this analysis, the Gradient Boosting method was employed to fit a predictive model. Gradient Boosting is an ensemble machine-learning technique that builds models sequentially to correct errors made by previous models. It combines multiple decision trees into a strong predictive model. At each stage, a new tree is trained to minimize the errors or residuals of the previous model using gradient descent, optimizing a specified loss function.

Table 1 displays the importance of each feature as derived from a decision tree model fitted using all features in the cleaned train data set. Initially, a Gradient Boosting model was built using the top 12 features with an importance score greater than 6. The model's hyperparameters, including the number of trees B, the shrinkage parameter λ, and the depth of splits d, were set to appropriate fixed values. To prevent overfitting, we performed 5-fold cross-validation to determine the optimal number of trees. However, when evaluating the model's predictive performance on the test data set, the results indicated suboptimal performance.



```
                      ADL                     MMSE    FunctionalAssessment        MemoryComplaints
              153.0481205              136.6306524             123.4410577              85.2724970
       BehavioralProblems         PhysicalActivity             DietQuality                     Age
               83.7054866                9.7476192               8.4057352               7.0508962
CholesterolTriglycerides         CholesterolTotal                     BMI            SleepQuality
                6.6848395                6.5778353               6.2762269               6.2179314
           CholesterolHDL       AlcoholConsumption          CholesterolLDL             DiastolicBP
                4.5193703                3.8298569               3.7129047               2.7025185
    CardiovascularDisease                Ethnicity              SystolicBP            Hypertension
                2.7002943                2.5522227               1.3520468               0.9644927
```

Table 1: Variable Importance Using Decision Tree

To enhance prediction accuracy, we adopted a hyperparameter tuning approach using a grid search under 5-fold cross-validation. This grid included four critical hyperparameters that significantly affect model performance: the number of trees, the shrinkage parameter, tree depth, and the minimum number of samples required in a tree's leaf nodes. For each hyperparameter, multiple reasonable values were explored to identify the optimal configuration.

The final tuned Gradient Boosting model was trained with the best combination of hyperparameters, determined to be 1500 trees, a tree depth of 2, a minimum of 15 samples in each leaf node, and a shrinkage parameter of 0.01. This configuration yielded a model with improved prediction performance, striking a balance between model complexity and accuracy.

During this research, we explored various statistical models, including logistic regression, Support Vector Machines (SVM), Quadratic Discriminant Analysis (QDA), decision trees, and Random Forests, to evaluate their applicability to the Alzheimer's Disease dataset. Logistic regression served as a baseline model due to its simplicity and interpretability but struggled with capturing non-linear relationships. SVM demonstrated strong performance in high-dimensional data, excelling in classification tasks like Alzheimer's diagnosis, though it required careful tuning and was computationally expensive (Burges, 1998). Random Forests offered robust feature importance rankings and reduced overfitting through ensemble learning, making them a valuable exploratory tool (Breiman, 2001). These analyses informed the development of our final Gradient Boosting model, which effectively balanced complexity and predictive performance by iteratively minimizing errors.

Model Evaluation and Statistical Inference

We then aim to evalulate the models' quality utilizing several metrics. These metrics include cross-validation, a confusion matrix, the ROC curve along with its AUC, and a Calibration plot. First, to evaluate the predictive performance of the final tuned Gradient Boosting model. We conducted a 5-fold cross-validation using the optimal hyperparameters. The results indicate that the final model achieves an accuracy of **95.88%**, demonstrating its robustness and reliability in predicting outcomes.

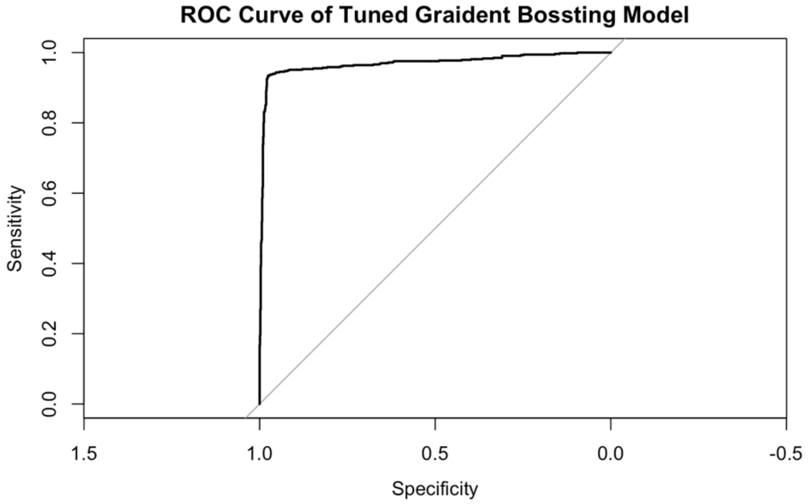


Figure 4: ROC curve of tuned GBM

The ROC curve demonstrates the high predictive accuracy of the tuned Gradient Boosting model on the cleaned training dataset, with an Area Under the Curve (AUC) value of **0.9692**. This indicates that the model has a strong capability to effectively differentiate between the two classes of the response variable, Diagnosis. A higher AUC reflects the model's superior performance in balancing sensitivity and specificity.

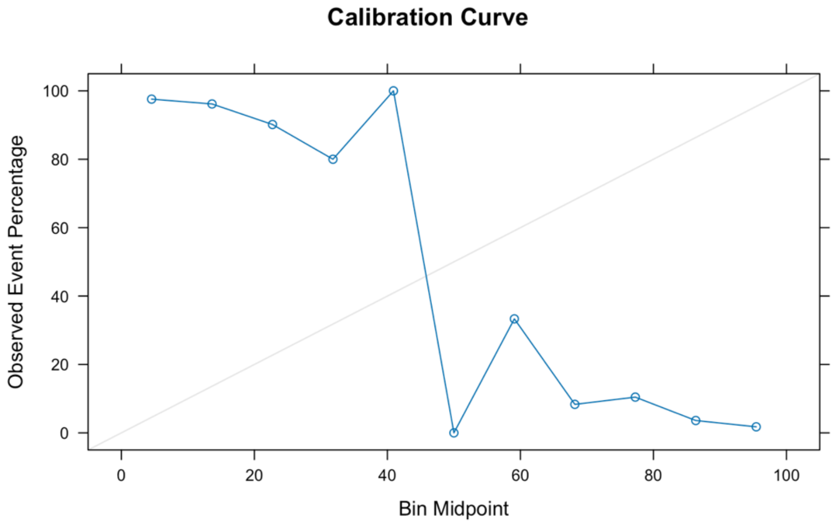


Figure 5: Calibration Plot of tuned GBM

The model performs well for lower predicted probabilities, aligning observed event rates with the predicted probabilities. However, at higher predicted probabilities (e.g., 60–100%), the observed event rates drop significantly, suggesting the model may be overconfident in its predictions within this range. This miscalibration indicates that while the model accurately ranks predictions, it may require recalibration techniques, such as Platt scaling or isotonic regression, to improve the reliability of probability estimates at higher thresholds.

Diagnosis Prediction	0	1
0	949	35
1	23	497

Table 2: Confusion Matrix

95% Confidence Interval	(0.9504, 0.9706)
No Information Rate	0.6463
P-Value [Accuracy > NIR]	<2e-16
Kappa	0.9152

Table 3: Relevant Statistics

According to the confusion matrix, the prediction accuracy can be calculated using the formula (TP + TN) / total instances, which equals (497 + 949) / (949 + 35 + 23 + 497) = 0.9614. This means the model can correctly classify **96.14%** of all instances. Moreover, the 95% Confidence Interval of the accuracy generated by R is (0.9504, 0.9706) in which the true accuracy of the model is likely to lie. The high prediction accuracy on the cleaned train data set shows a good model performance to some extent.

To test if the model's accuracy is significantly better than the No Information Rate, set confidence level α = 0.05. Given that the p-value is very small (< 0.05), the model's accuracy is statistically and significantly better than guessing the majority class. Kappa measures the agreement between the predicted and actual classes, adjusting for agreement by chance. A Kappa value of **0.9152** indicates excellent agreement between predictions and actual values.

Compared to the other models we evaluated, the logistic regression models generated using LASSO, stepwise AIC, and BIC produced notably lower AUC values (0.9047, 0.9049, and 0.9047, respectively). The

Discriminant Analysis model also performed poorly on the test dataset, failing to accurately predict AD diagnoses. While the basic decision tree model achieved a 0.9574 accuracy rate according to its confusion matrix, this was still smaller than the Gradient Boosting model's 0.9614 accuracy. Similarly, the SVM and Random Forest models demonstrated relatively strong performance on the test data, their accuracy remained less impressive than that of the Gradient Boosting model.

Given the above metrics and statistics, it is reasonable to conclude that Gradient Boosting model is an appropriate statistical model to predict the Diagnosis of Alzheimer's Disease.

Results & Conclusion

Results

The predictive modeling process yielded a Gradient Boosting model that demonstrated strong performance in diagnosing Alzheimer's Disease. In our data analysis process, we first identified key predictors with feature selection. Through LASSO, AIC/BIC stepwise selection, and decision tree analysis, five core predictors were identified as highly influential: MMSE**,** a well-established cognitive assessment tool; Functional Assessment, evaluating patients' abilities in daily tasks; Memory Complaints, a self-reported symptom of cognitive decline; Behavioral Problems, indicators of changes in personality or behavior; ADL**,** assessing functional independence. These features collectively provide a comprehensive view of cognitive, behavioral, and functional dimensions of Alzheimer's Disease, aligning with many existing literature on the early predictors of AD. These predictors are highly associated with the outcome variable and play a crucial role in early diagnosis, aligning with the objectives of this investigation.

Furthermore, we chose the Gradient Boosting model that achieved an accuracy of **95.88%** on the cleaned training dataset. After evaluating multiple models, the tuned Gradient Boosting model was preferred with an AUC value of **0.9692**, indicating excellent discriminatory power between Alzheimer's positive and negative cases. Moreover, the model obtained a prediction accuracy of **96.14%**, with a high agreement between predicted and actual values (Kappa = 0.9152) with the confusion matrix. Logistic regression models, while interpretable, struggled to capture non-linear relationships in the data, limiting their predictive accuracy. Other models such as SVM and random forest were lower in their predictive ability of the test data in the competition or obtained a lower AUC value in comparison. After evaluating multiple models on the given half of the test data set, we selected a tuned Gradient Boosting model that achieved a **94.894%** accuracy - the highest among all models tested. When applied to the other half of the test set, this model maintained a strong performance with a **91.987%** accuracy. These results indicate that, although there may still be room for further optimization, we have developed a robust predictive model that generalizes well, as evidenced by its high and similar accuracy on both training data and test data.

Conclusion

We successfully explored the factors most strongly associated with Alzheimer's Disease by identifying 5 key predictors among the 32 given. These variables, consistently highlighted through feature selection methods like LASSO and decision tree analysis, provide a comprehensive understanding of the cognitive, behavioral, and functional dimensions linked to Alzheimer's diagnosis, aligning with our objective to uncover significant contributing factors.

We have further achieved the objective of building an ML model using the training data provided to fairly accurately predict the diagnosis of Alzheimer's Disease. The Gradient Boosting model demonstrated strong predictive performance, supported by high accuracy and AUC metrics. The identified predictors are also well-established in the literature, reaffirming their diagnostic relevance. Although the accuracy of the predictive outcomes could be further improved, the findings demonstrate the potential of machine learning as a decision-support tool in healthcare, enabling earlier and more accurate diagnoses. This could facilitate timely interventions, improve patient outcomes, and reduce caregiver burdens for many in the population affected by Alzheimer's Disease today.

Discussion

Limitations

Miscalibration at Higher Predicted Probabilities

The calibration plot shown as Figure 5 indicated that the final GBM model tends to overpredict for higher probability predictions. While the model performed well overall, its confidence levels did not fully align with observed event rates. This miscalibration could have implications in real-world applications, such as prioritizing patients for further diagnostics or treatment interventions, where precise probability estimates are crucial. Such discrepancies between predicted and actual probabilities highlight the need for recalibration techniques to improve the model's reliability in clinical decision-making settings (Barrett-Connor et al., 2008).

Potential Overfitting

Although cross-validation was employed to minimize overfitting, the high accuracy and AUC observed on the training dataset raise concerns about whether the model captured noise or overly specific patterns in the data. Overfitting can result in reduced generalizability, particularly when the model is applied to new, independent datasets. This limitation underscores the importance of testing the model on real-world data to evaluate its robustness and applicability across diverse contexts.

Interpretability Challenges

Despite the identification of key features such as MMSE, ADL, and Memory Complaints through variable importance analysis, GBM remains a black-box model compared to interpretable alternatives like logistic regression. In high-stakes applications, such as Alzheimer’s diagnosis, the lack of explainability may limit adoption due to the need for transparency and trust in model predictions. Decision trees, though less complex, provide clearer hierarchical relationships among predictors, which can be more actionable for healthcare professionals (Breiman et al., 2019).

Future Directions

Future research could explore recalibration techniques, such as isotonic regression or Platt scaling, to improve the alignment between predicted probabilities and observed event rates, thereby enhancing the reliability of probabilistic predictions, particularly in addressing model complexity concerns such as overfitting and improving overall model accuracy.

Additionally, while the GBM model demonstrated strong performance, alternative algorithms like Random Forests and Support Vector Machines (SVM) warrant consideration in future studies. Random Forests offer robust performance with resistance to overfitting while maintaining interpretability (Breiman, 2001). Similarly, SVM has proven effective in handling high-dimensional datasets, making them a promising candidate for Alzheimer’s Disease prediction (Burges, 1998). These approaches could provide valuable insights into optimizing both predictive accuracy and computational efficiency.

Future research could also focus on validating the model with external datasets to ensure generalizability across diverse populations. The current model has been trained and evaluated solely on the provided dataset, which, while comprehensive, may not capture the full diversity of real-world populations. Alzheimer’s Disease diagnosis is influenced by a wide range of factors, including genetic, lifestyle, and environmental variables, which may vary across different demographics, geographic regions, and healthcare systems.

Moving forward, integrating explainable AI techniques and collaborating with diverse healthcare datasets can further enhance the reliability, transparency, and clinical adoption of such models. This research underscores the transformative role of data-driven solutions in addressing global healthcare challenges like Alzheimer’s Disease.

References

Alzheimer’s Disease International. (2022). *World Alzheimer Report 2022*. https://www.alzint.org/resource/world-alzheimer-report-2022/

Barrett-Connor, E., Dam, T. T., Stone, K., Harrison, S. L., Redline, S., Orwoll, E., & Osteoporotic Fractures in Men Study Group. (2008, July). The association of testosterone levels with overall sleep quality, sleep architecture, and sleep-disordered breathing. *The Journal of Clinical Endocrinology & Metabolism, 93*(7), 2602–2609. https://doi.org/10.1210/jc.2007-2622

Breiman, L. (2001, October). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2019). *Hands-On Machine Learning with R (1st ed.)*. Chapman and Hall/CRC. https://doi.org/10.1201/9780367816377

Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*, 121-167. https://doi.org/10.1023/A:1009715923555

Elflein, J. (2024, November 8). *Dementia and Alzheimer's disease in Canada - Statistics & Facts*. Statista. https://www.statista.com/topics/11570/dementia-and-alzheimer-s-disease-in-canada

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975, November). “Mini-Mental State”. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189-198. https://doi.org/10.1016/0022-3956(75)90026-6

Guerreiro, R., & Bras, J. (2015). The age factor in Alzheimer’s disease. *Genome Medicine*, *7*(106). https://doi.org/10.1186/s13073-015-0232-5

Stefaniak, O., Dobrzyńska, M., Drzymała-Czyż, S., & Przysławski, J. (2022). Diet in the prevention of Alzheimer's disease: Current knowledge and future research requirements. *Nutrients, 14*(21), 4564. https://doi.org/10.3390/nu14214564

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178