Executive Summary: Predicting Alzheimer's Disease

Project by: Mercy Amankwah, Seyed Abdolhamid Banihashemi, Nasheed Jafri
GitHub: https://github.com/mgamankwah/alzheimers_risk_prediction/

Introduction

Alzheimer's disease (AD) is a progressive disorder impacting memory, thinking, and behavior. Early detection is vital but challenging due to its complex causes and gradual onset. This project aims to build a predictive model to estimate AD risk using features such as demographics, medical history, lifestyle, clinical data, and cognitive assessments.

Dataset

We used a synthetic dataset from Kaggle's Alzheimer's Disease Risk Prediction competition, containing 2,149 patient records (ages 60–90) and 34 features across six categories: medical history, lifestyle, clinical data, demographics, cognitive assessments, and symptoms. The binary target variable is Diagnosis. The dataset was clean (no missing values or outliers) and pre-split into training (with labels) and test sets (without labels).

Model Selection, Tuning, and Feature Importance Analysis

We split the Kaggle training set 80/20 for validation. Multiple classifiers were trained on the training set, tested on the validation set, and compared to a baseline (majority class). To explore feature importance, models were also run on different subsets of features. For early detection, we repeated the process on patients without memory complaints. Using accuracy as the metric for initial model selection, AdaBoost (96.5%) performed best on the full dataset, while XGBoost (95.6%) led in the early detection group. We tuned both models with Optuna, optimizing metrics (accuracy, F1, recall, and precision) to penalize different sources of inaccuracy, using both 5-fold cross-validation and full training set fitting. Performance was evaluated via confusion matrices on the validation set and F1 score on the Kaggle test set via submission to the competition page. A tuned AdaBoost achieved an F1 of 93.28%, close to the leaderboard's top score of 94%. Post-tuning, validation accuracy improved to **97%** for **AdaBoost** and **96.7%** for **XGBoost** in the early detection scenario.

For each of the tuned best-performing models, we performed a feature importance analysis. Across both models, **cognitive assessment features** like **MMSE, Functional Assessment, and ADL** are the most influential in predicting Alzheimer's.

Probabilistic Modeling

In a medical context, expressing a patient's probability of developing Alzheimer's is more meaningful than a simple binary classification. Therefore, we calibrated our models to output risk probabilities and categorized patients into low, medium, and high-risk groups. We applied two calibration methods (isotonic and sigmoid) and compared their performance using Brier scores and ROC AUC. **Isotonic calibration** outperformed sigmoid on both metrics. As expected, the calibrated model correctly assigned most non-Alzheimer's patients to the low-risk group and most Alzheimer's patients to the high-risk group.

To gain further insight into feature impact, we conducted an inferential analysis by building a model to estimate the probability of a positive diagnosis conditioned on key features. Comparing empirical Alzheimer's risk to model-predicted risk revealed strong alignment, with better feature scores consistently associated with lower predicted risk.