Erdös Data Science Bootcamp

Spring 2025

PREDICTING ALZHEIMER'S DISEASE

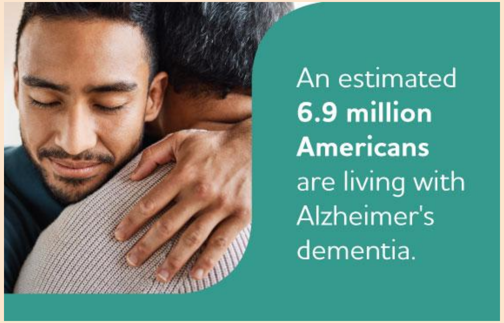
Project by

Mercy Amankwah
Seyed Abdolhamid Banihashemi
Nasheed Jafri

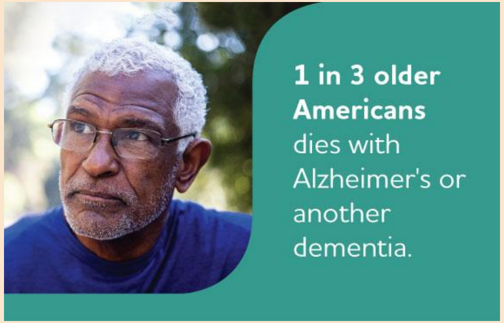
Mentor

Cemile Kurkoglu

Background



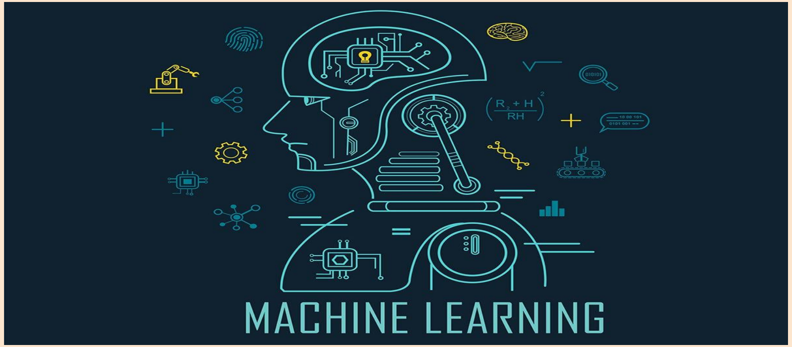
Prevalence (in 2024)



Not just memory loss, Alzheimer kills

~ $1 trillion

Health and long-term care costs for people living with dementia was projected to reach $360 billion in 2024 and nearly $1 trillion in 2050.



GOAL: Predict A.D. using M.L. models

Dataset

2149

PATIENT RECORDS

6

FEATURE SUB-CATEGORIES

34

FEATURES

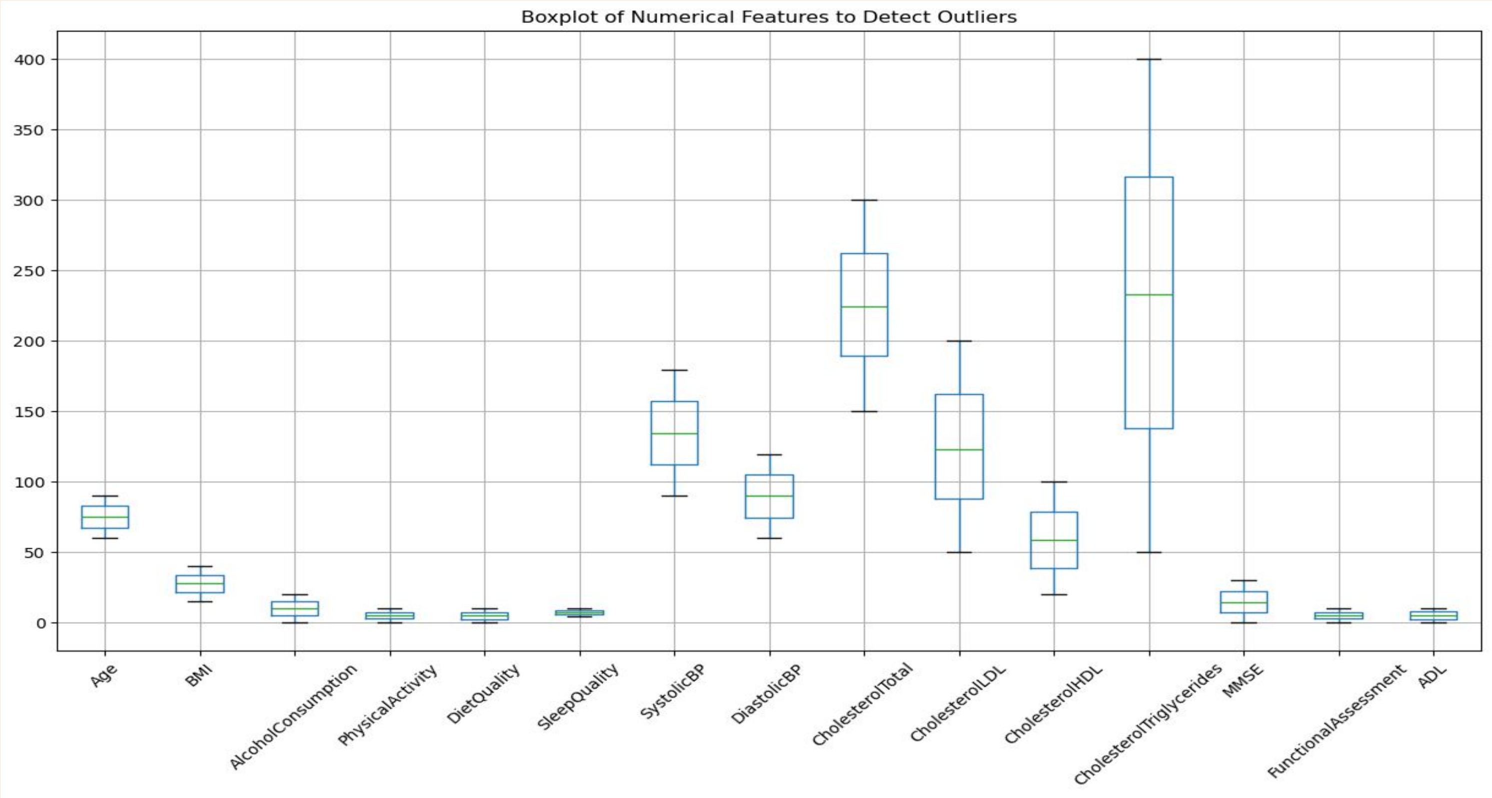
AGE RANGE 60–90

TARGET VARIABLE DIAGNOSIS

DATASET IS CLEAN

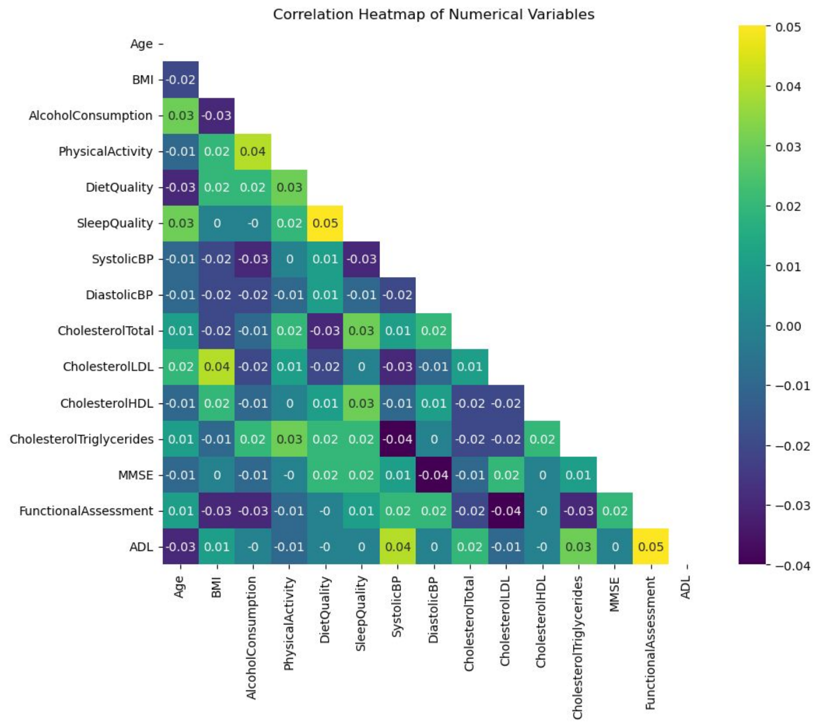
Medical History	e.g. *Family History of Alzheimer's, Diabetes*
Lifestyle Factors	e.g. BMI, Smoking, Alcohol Consumption
Clinical Measurements	e.g. Systolic BP, Cholesterol Total
Demographics	e.g. Age, gender, Ethnicity
Cognitive Assessments	e.g. Functional Assessment, Memory Complaints
Symptoms	e.g. Confusion, Disorientation, Forgetfulness

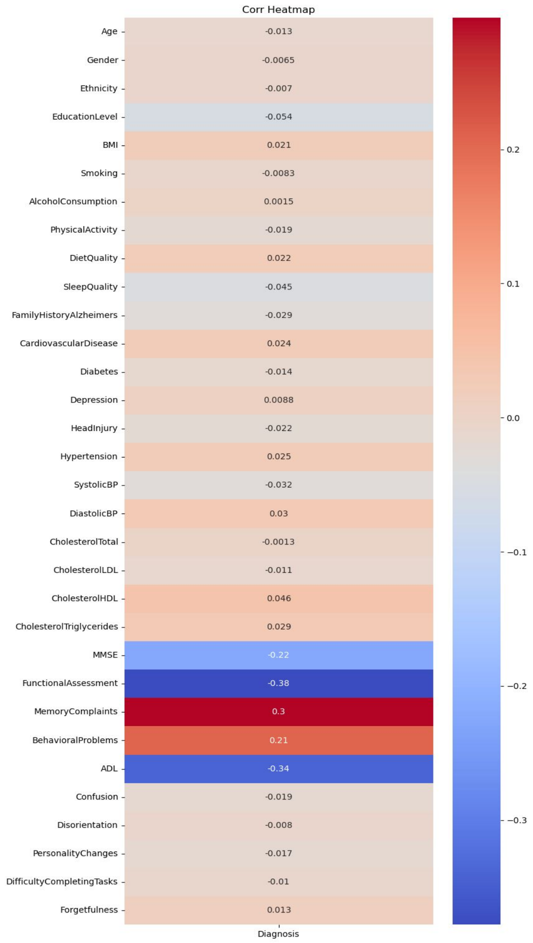
Exploratory Data Analysis (Outlier Detection)



The dataset has no significant outliers.

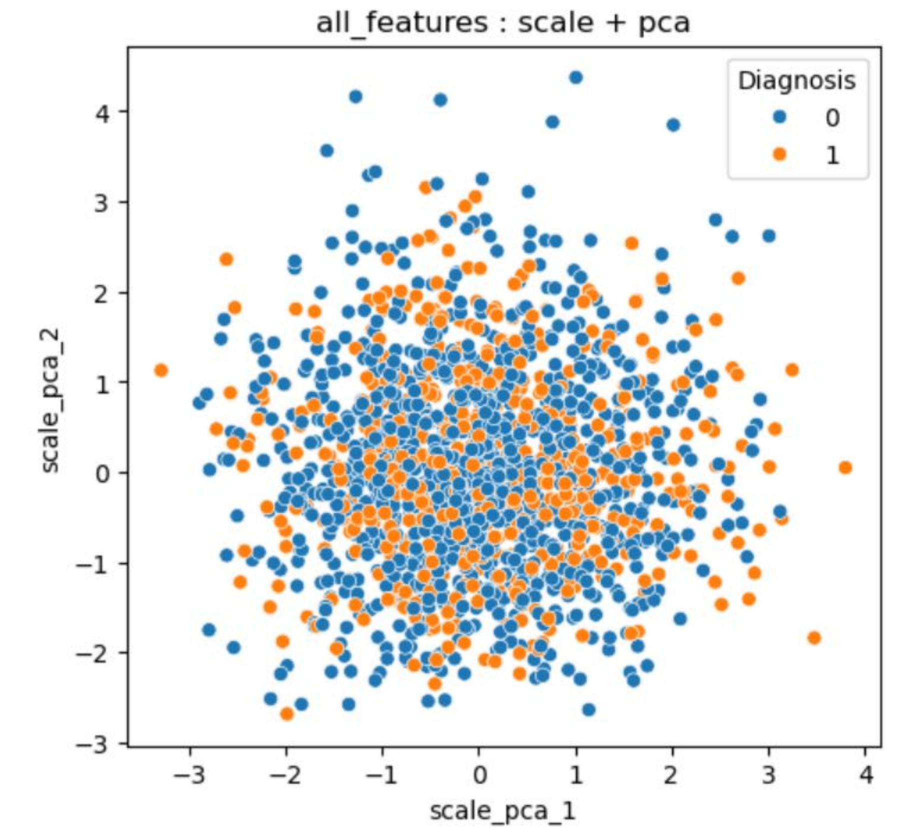
Exploratory Data Analysis (Correlation Analysis)

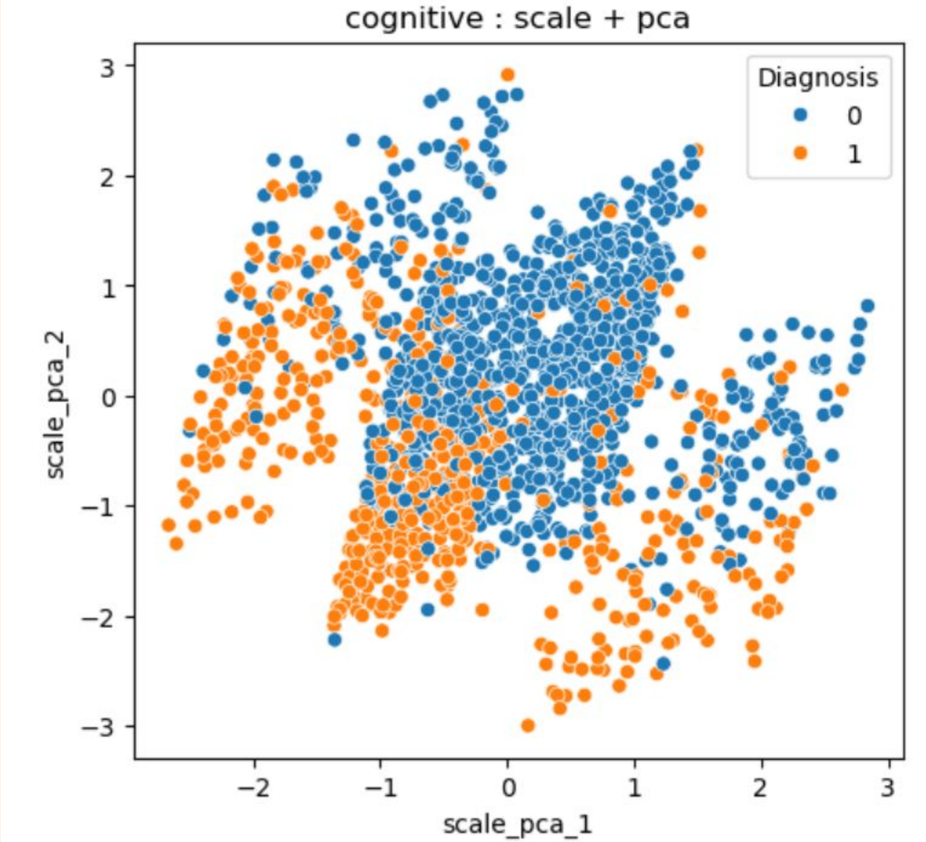




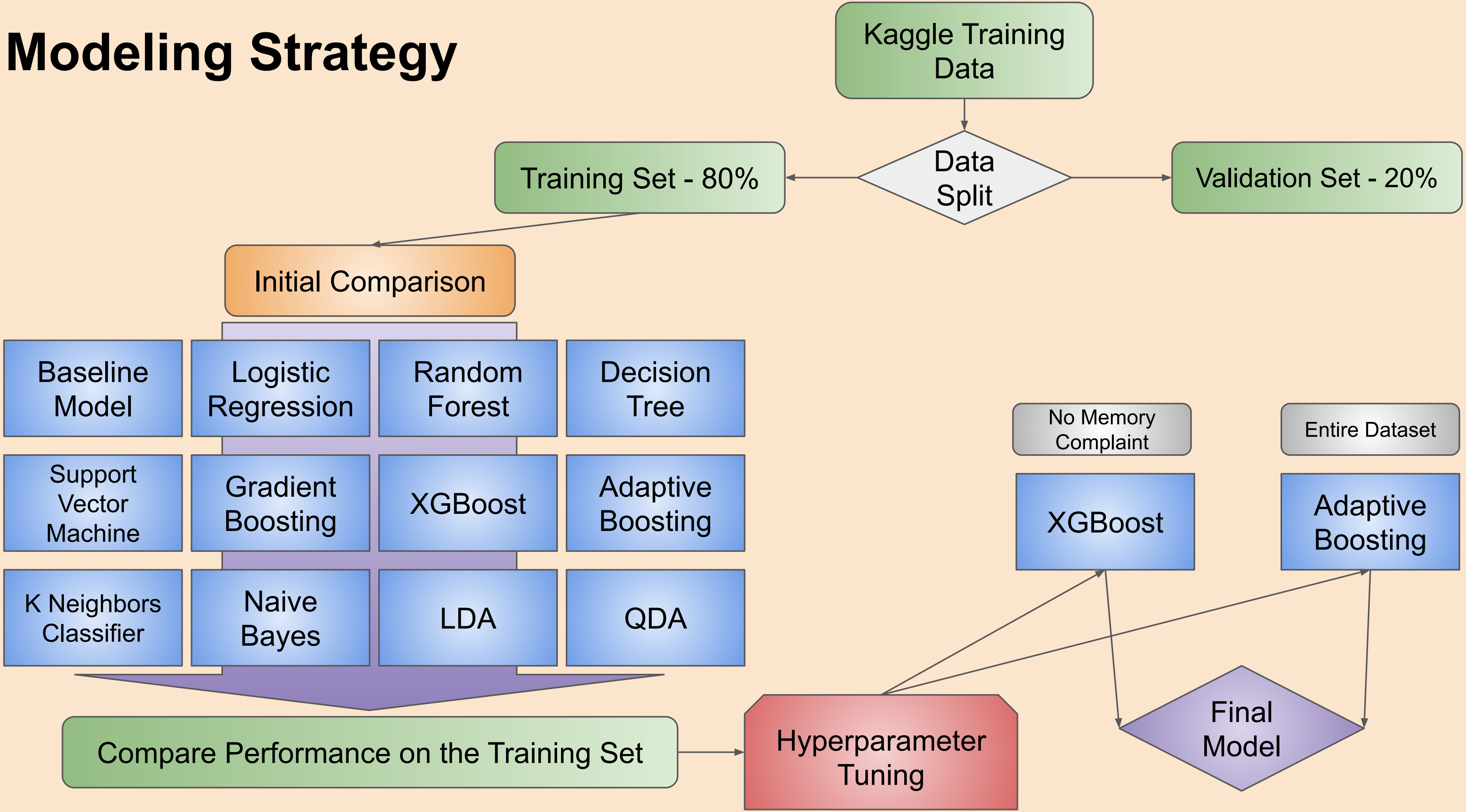
- The target variable has the **highest correlation with the cognitive features**.
- No significant correlation within numerical predictors.

A Simple 2D PCA Model



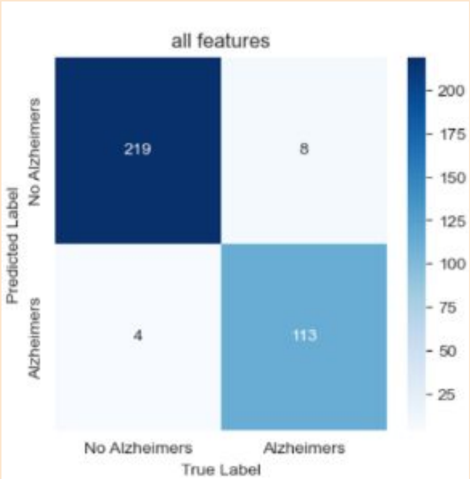


Modeling Strategy

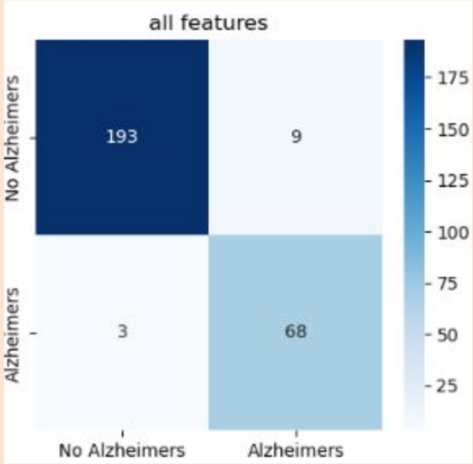


Model Comparison

MODEL	Accuracy (Entire Dataset)	Accuracy (No Memory Complaints)
Logistic regression	0.8605	0.85348
Random Forest	0.936	0.90476
Gradient boosting	0.9506	0.94872
XGBoost	0.9506	**0.95604**
Adaboost	**0.9651**	0.94506
KNN	0.7297	0.76923
SVM (Poly Kernel)	0.7820	0.83883
Naïve Bayes	0.8314	0.85714
LDA	0.843	0.85348
QDA	0.8052	0.64103



Confusion Matrix for AdaBoost on Entire Dataset



Confusion Matrix for XGBoost on Restricted Data

Model Tuning

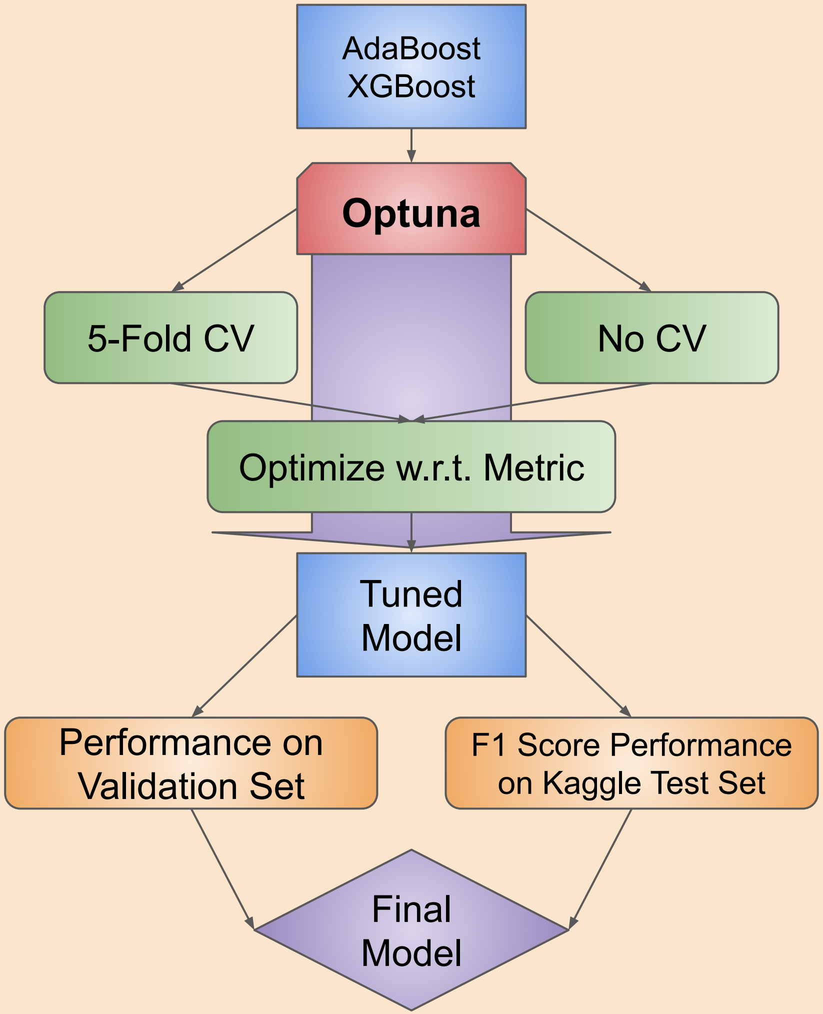
Tuned Hyperparameters:

- AdaBoost:
  n_estimators, learning_rate, algorithm, criterion, max_depth, min_sample_split

- XGBoost:
  n_estimators, learning_rate, max_depth, min_child_weight, gamma, subsample, colsample_bytree, eval_metric, reg_alpha, reg_lambda, scale_pos_weight

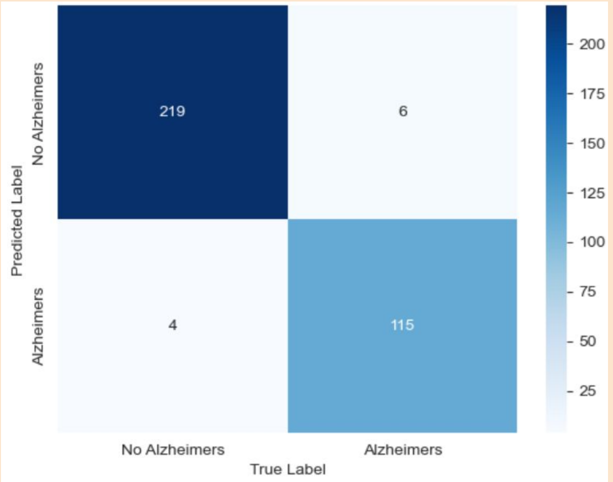
Metrics Optimized:

Accuracy, Precision, Recall, F1

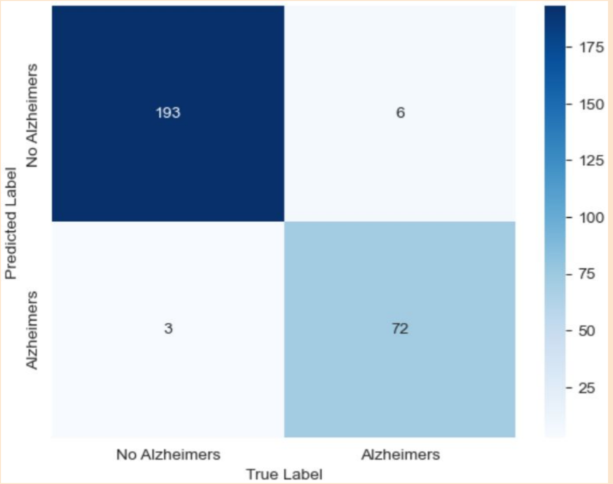


Performance of the Tuned Models:

- The tuned models slightly out-performed the initial ones.
- The best F1 score from Kaggle competition was **0.93288**, for a tuned **AdaBoost** model **without CV**.
- For entire dataset, **AdaBoost** still outperformed XGBoost after tuning **without CV**.
- For dataset restricted to patients with no memory complaints, **XGBoost** outperformed AdaBoost after tuning over **5-fold CV**.

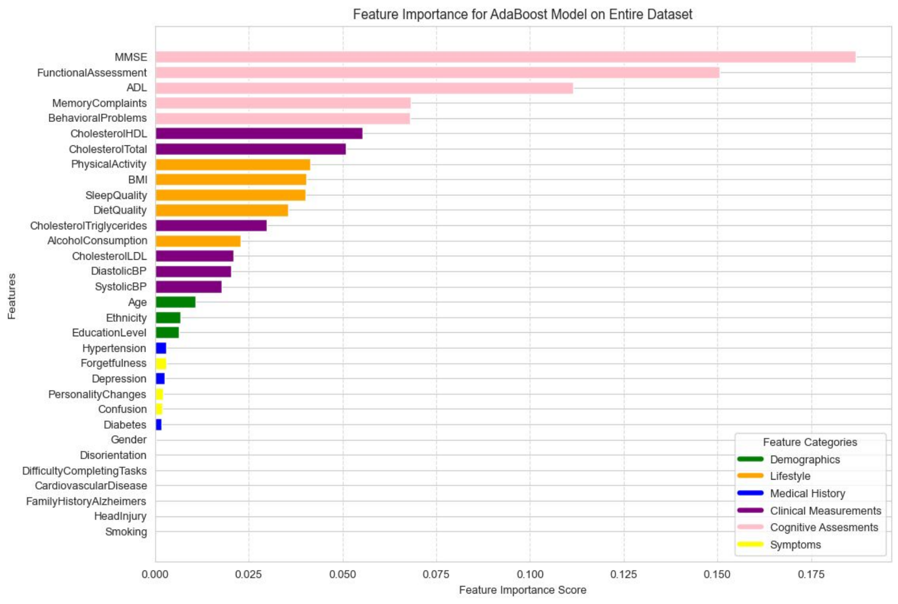


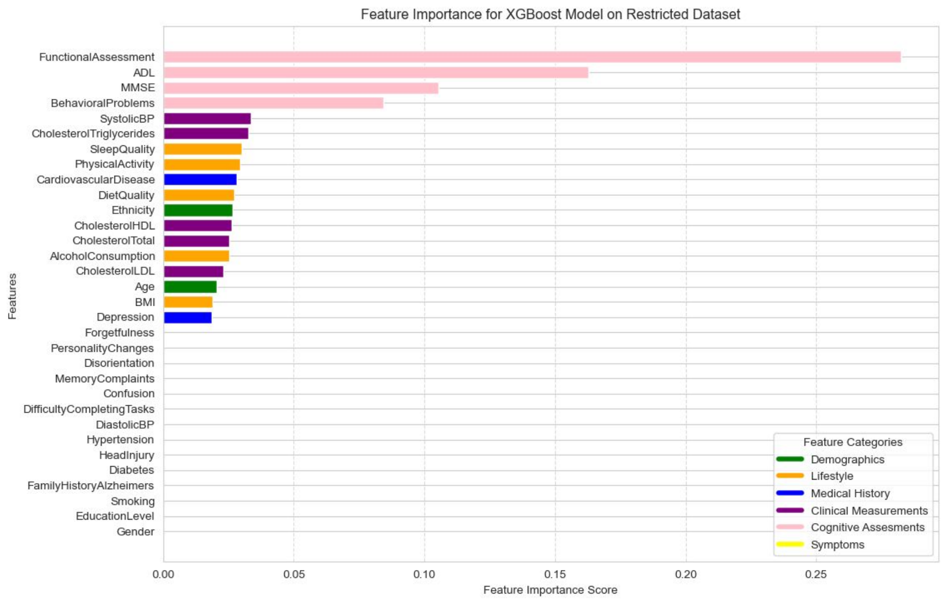
Confusion Matrix for AdaBoost on Entire Dataset



Confusion Matrix for XGBoost on Restricted Data

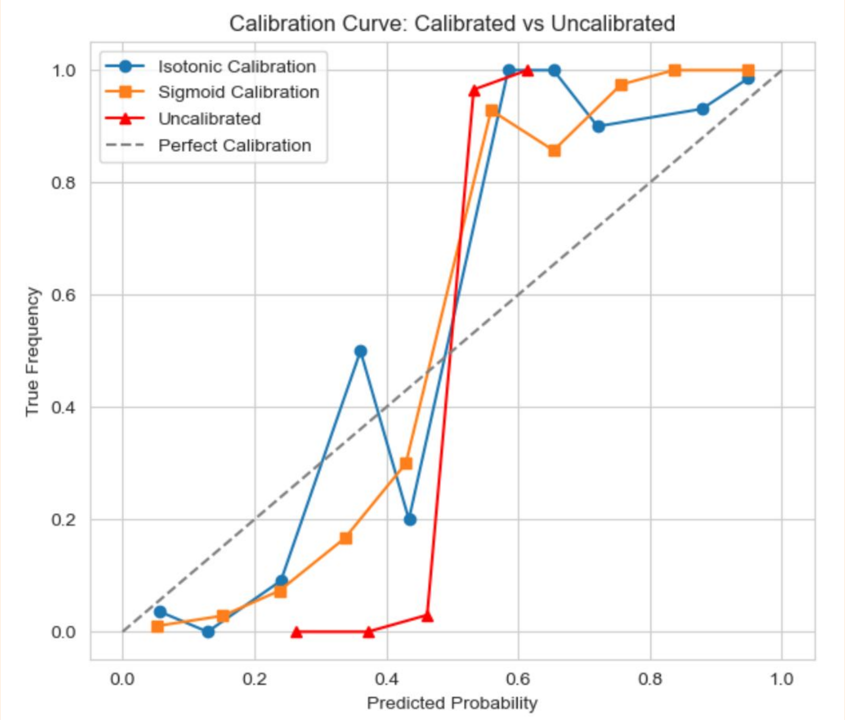
Feature Importance

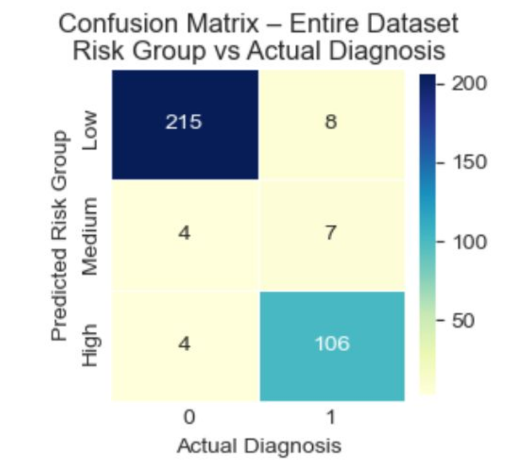


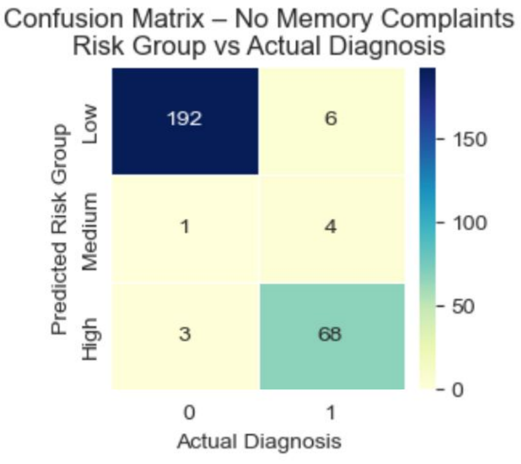


- Cognitive and assessment features like MMSE, Functional Assessment, ADL, Memory Complaints and Behavioral Problems rank highest in feature importance.
- Other clinical measurements like cholesterol and blood pressure also seem important.

Probabilistic Model for Risk Prediction







- Isotonic performed better at improving the number of correct classifications of the uncalibrated model.

Insights for important features

🧠 Conditional Alzheimer's Risk Given:

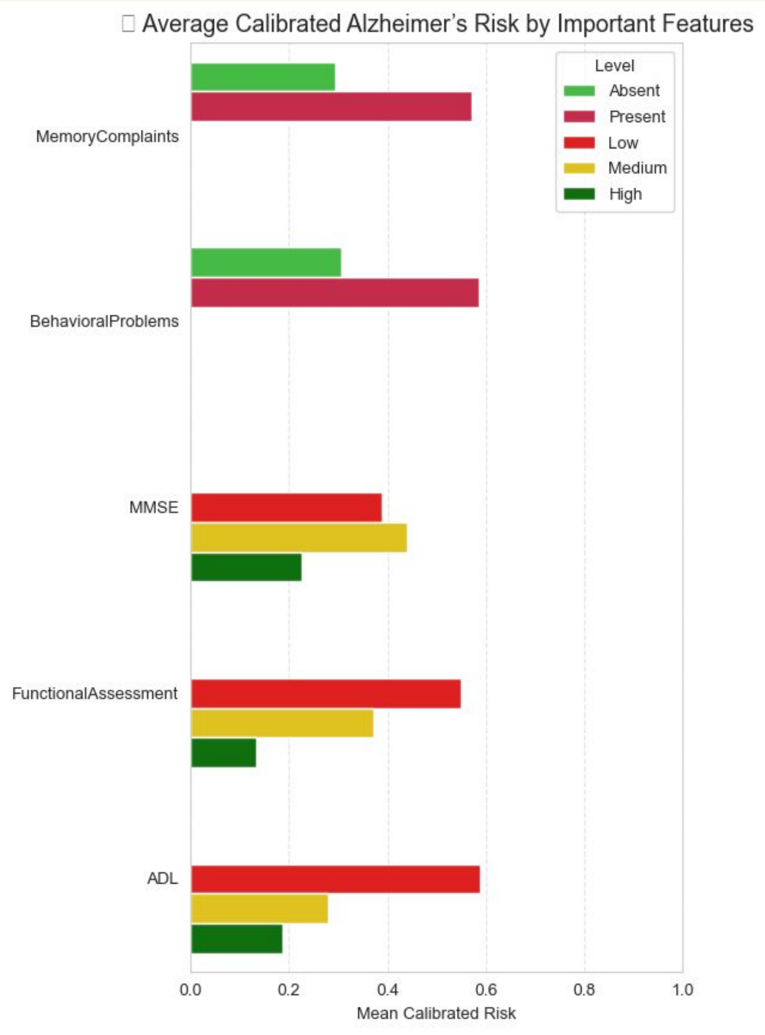
Memory Complaints

Diagnosis	0	1
MemoryComplaints		
0	71.5%	28.5%
1	38.6%	61.4%

Behavioral Problems

Diagnosis	0	1
BehavioralProblems		
0	70.2%	29.8%
1	36.4%	63.6%

- Presence of memory complaint and behavioral problems is associated with higher risk of Alzheimer’s
- Higher MMSE, Functional Assessment and ADL scores are associated with lower risk of Alzheimer’s
- Lower scores indicate greater impairment.



Empirical vs Calibrated Risks

- We group the important assessment scores as Low, Medium, High values and compare the empirical risk with the calibrated Alzheimer's risk for our model.

📝 Emperical and Modeled Probabilities for Low, Medium and High scores

Mini-Mental State Exam

MMSE	Empirical	Model
Low	41.7%	38.9%
Medium	44.7%	43.4%
High	19.1%	21.6%

Functional Assessment

FA	Empirical	Model
Low	55.7%	56.5%
Medium	36.8%	37.5%
High	13.0%	9.8%

Activities of Daily Living

ADL	Empirical	Model
Low	59.1%	58.9%
Medium	28.9%	27.8%
High	17.4%	17.0%

- Patients with high MMSE scores are at lower risk.
- Alzheimer's risk is inversely related to functional assessment and ADL (activities of daily life).
- The risk probabilities from our model fairly aligns with the empirical ones.

Conclusion

- ❖ Predicted Alzheimer diagnosis under 2 scenarios
- ❖ Adaboost performed best with accuracy of 97% and F1 score 96% (all patients)
- ❖ XGBoost performed best with accuracy of 96% and F1 score 94% (no memory complaints)
- ❖ Cognitive features were the most important features in prediction
- ❖ Patients who report memory complaints had more than double the Alzheimer’s risk (61.4%) compared to those who don’t (28.5%)
- ❖ Presence of behavioral problems also approximately doubled the probability of Alzheimer's diagnosis (63.6% vs 29.8%)

Next Steps

- ❖ Analyzing and modeling risk predictions based on real-world databases
- ❖ Creating a user-friendly program for alzheimer’s risk assessment

Thank you!

Mentor: Cemile Kurkoglu

Instructor: Steven Gubkin

All TA's

Entire Erdös Team