Classifying Dementia Using the Open Access Series of Imaging Studies (OASIS) Longitudinal Dataset

Sarah Bao
Brown Department of Mathematics & Computer Science
December 11, 2024

https://github.com/s-bao/data1030-project.git

Introduction

Problem Statement:

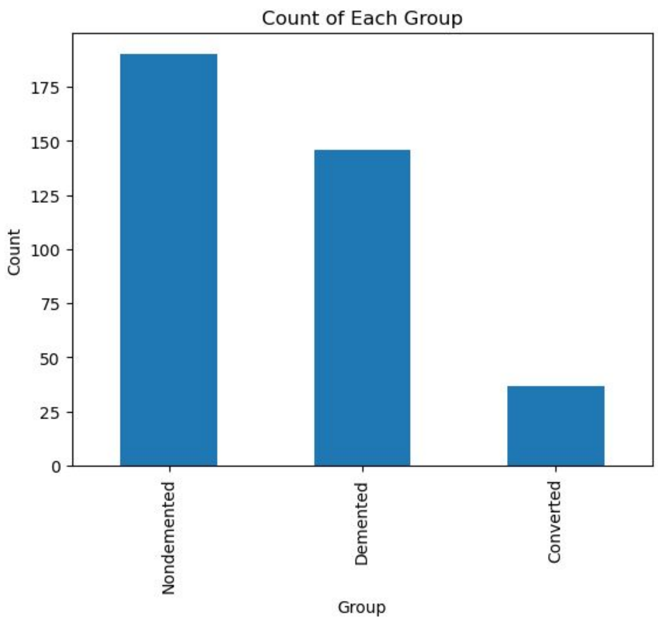
- aim to classify individuals as either "Demented", "Nondemented", or “Converted” based on features such as age, cognitive test scores, and brain volume measurements
- importance: early detection of dementia is vital for timely intervention and management, potentially improving patients' quality of life
    - doctors make clinical diagnoses based off comprehensive assessment of patient's condition, since a definitive diagnosis can only be made after an autopsy of the brain
    - model can assist clinicians in identifying high-risk individuals

Dataset Description:

- Kaggle MRI and Alzheimer's (https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers/data)
- OASIS longitudinal dataset: contains information on 150 subjects ages 60 to 96; longitudinal because it has information on at least 2 visits for each subjects, where visits are separated by at least 1 year, where all cognitive and MRI measurements are updated during each visit

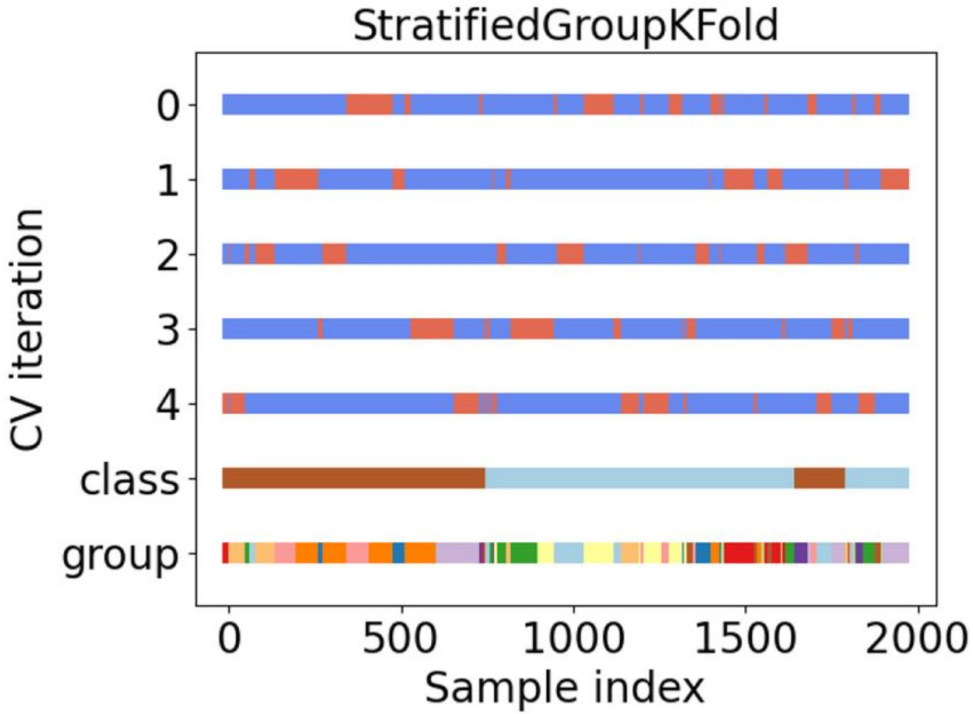
Qualities of Dataset

- 373 by 15
- Non-IID data since multiple data points provide information on the same person (150 subjects)
- Only 2 data points with missing values in continuous feature MMSE
- 15 features: Subject ID, MRI ID, Group, Visit, MR Delay, M/F, Hand, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF
    - Target variable: Group
    - MMSE, CDR (mini-mental state examination, clinical dementia rating) = cognitive test scores
    - eTIV, nWBV, ASF (estimated total volume in skull, normalized volume of whole brain, Atlas scaling factor) = MRI metrics



Splitting

- has group structure ➔ group split on "Subject ID"
    - samples are not iid because each subject appears multiple times in the dataset
- imbalanced ➔ stratified split on "Group" target variable
- relatively small (< 100k) data points ➔ kfold for cross validation
- **used StratifiedGroupKFold**



Pipeline

Handling missing values

- SES = ordinal ⇒ ordinal encoded missing values as -1
- MMSE = continuous ⇒ dataset had only 2 missing values, so decided to use multivariate imputation

Evaluation metric

- tested a few different ones including accuracy, f1 macro, and f1 weighted ⇒ ultimately chose f1 macro

```
Missing values in each column:
Subject ID     0
MRI ID         0
Group          0
Visit          0
MR Delay       0
M/F            0
Hand           0
Age            0
EDUC           0
SES           19
MMSE           2
CDR            0
eTIV           0
nWBV           0
ASF            0
```

Pipeline Summary

1. **Splitting method – StratifiedGroupKFold**
2. **Preprocessing – Normal + final Standard Scaler**
3. **Missing values handling method – Multivariate imputation**
4. **Evaluation metric – f1 macro**

ML Models

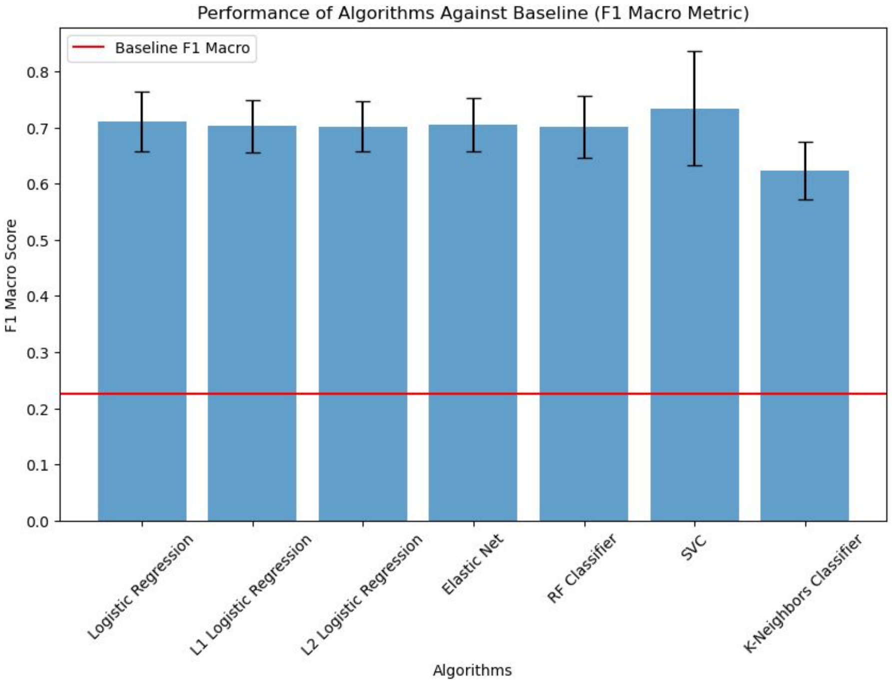
Summary of Models Tested

Model Name	Parameters Fitted
SimpleLogisticRegression	{}
L1LogisticRegression	{'C': [0.001, 0.01, 0.1, 1, 10, 100]}
L2LogisticRegression	{'C': [0.001, 0.01, 0.1, 1, 10, 100]}
ElasticNet	{'C': [0.001, 0.01, 0.1, 1, 10, 100], 'l1_ratio': [0.001, 0.01, 0.1, 1]}
RandomForestClassifier	{'n_estimators': [100], 'max_depth': [1, 3, 5, 10, 20, 100], 'max_features': [0.25, 0.5, 0.75, 1.0, None]}
SupportVectorClassifier	{'C': [1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3], 'gamma': [1e-5, 1e-3, 1e-1, 1e1, 1e3, 1e5]}
KNeighborsClassifier	{'n_neighbors': [3, 5, 10, 20], 'weights': ['uniform', 'distance']}

Results

Summary of Model Scores

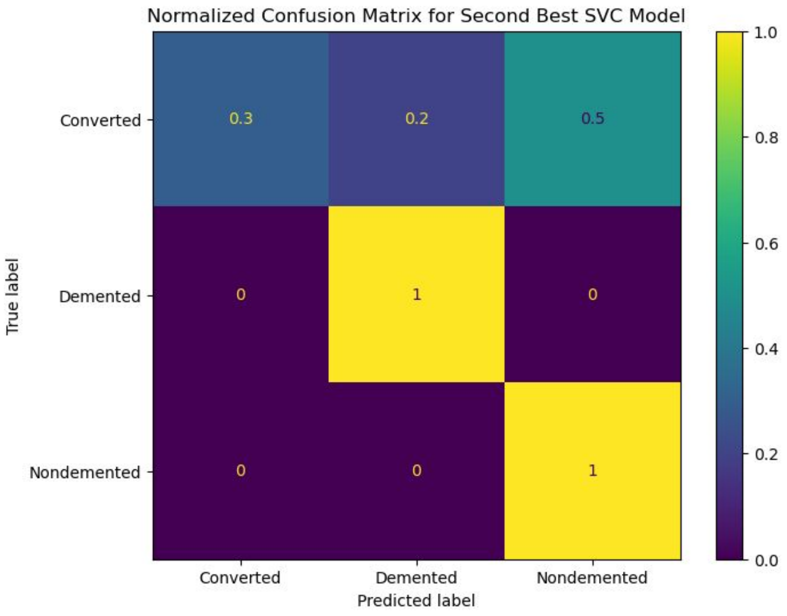
Algorithm	Mean Score	Std Dev
Logistic Regression	0.7103	0.0536
L1 Logistic Regression	0.7024	0.0471
L2 Logistic Regression	0.7013	0.0445
Elastic Net	0.7046	0.0472
RF Classifier	0.7013	0.055
SVC	0.7343	0.1021
K-Neighbors Classifier	0.6224	0.0513

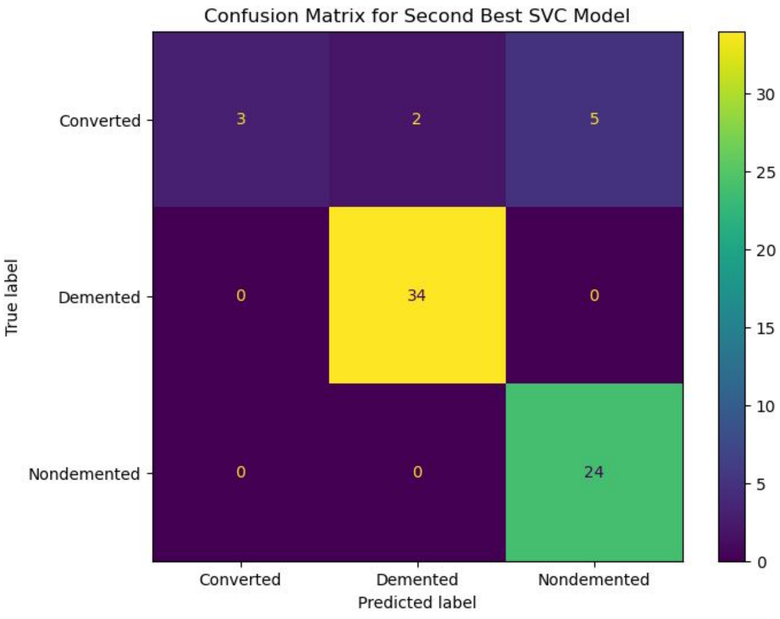


Best model (ish) ⇒ SVC

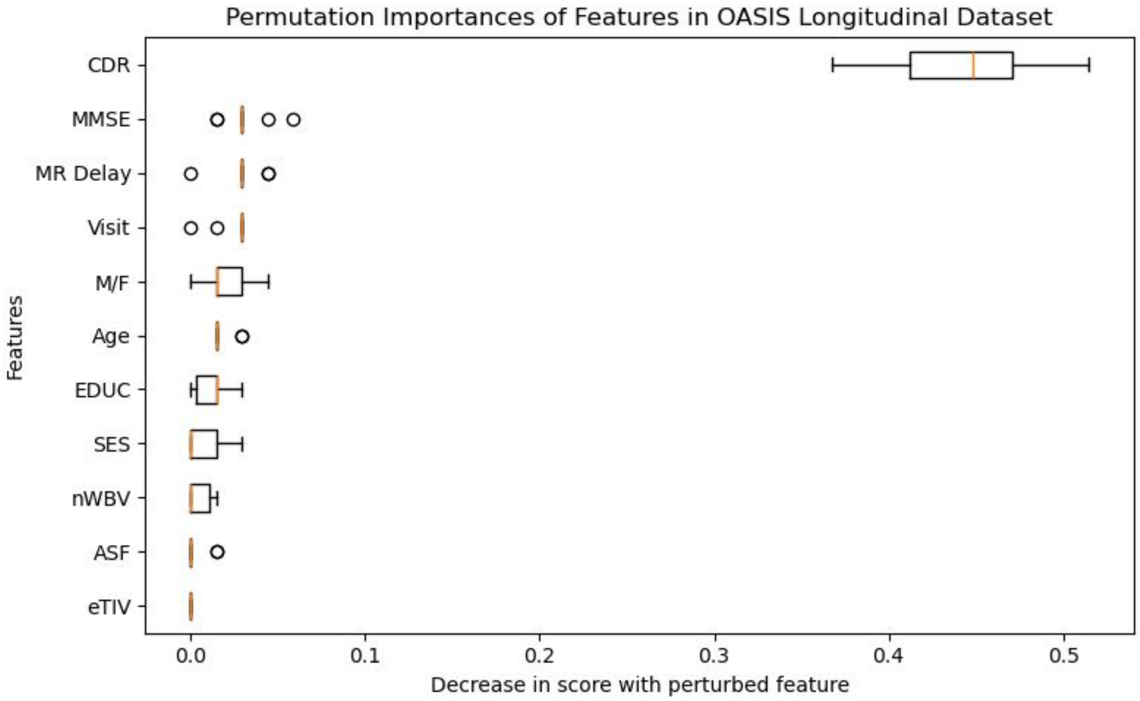
- 1st best had score of 0.9861
- 2nd best had score of 0.7795

SVC Interpretability: Confusion Matrix





SVC Interpretability: Feature Importance



Outlook

Predictive Power Enhancement

- Dataset has few features, could definitely consider adding more (but would need domain knowledge)

Improve Interpretability

- Apply SHAP for local interpretability test

Other Methods to Try

- Could try XGBoost and deep learning models
- Could try reduced features model to handle missing data rather than imputation

Thank you for listening!