

Capstone Project: Machine Learning Engineer Nanodegree –Modeling and Classification of OASIS MRI dataset for Dementia Levels

Project Overview

This capstone project involves machine learning modeling and analysis of clinical, demographic, and brain related derived anatomic measures from human MRI (magnetic resonance imaging) tests (<http://www.oasis-brains.org/>). The objectives of these measurements are to diagnose the level of Dementia in the individuals and the probability that these individuals may have Alzheimer's Disease (AD).

In published studies, Machine Learning has been applied to Alzheimer's/Dementia identification from MRI scans and related data in the academic papers/theses in References 10 and 11.

Recently, a close relative of mine had to undergo a sequence of MRI tests for cognition difficulties. The motivation for choosing this topic for the Capstone project arose from the desire to understand and analyze potential for Dementia and AD from MRI related data. This Capstone project does not use the MRI "imaging" data and does not focus on AD, focusses only on Dementia.

Project Analysis Files

Appendix-1 at the end of this report contains the name of the files that are submitted to Udacity for review.

Problem Statement

- Cross-sectional and longitudinal OASIS MRI structural and demographic data (clinical, demographic, and brain related derived anatomic measures) from human MRI (magnetic resonance imaging) tests (<http://www.oasis-brains.org/>) will be used to train a set of linear and non-linear machine learning classification models.

- A dataset that combines the cross-sectional and the longitudinal MRI datasets will be used. This has the benefit of having a larger dataset. Some of the column headings of the longitudinal dataset is to be renamed to match that of the cross-sectional MRI dataset
- Clinical Dementia Rating (CDR) values provided in the data set will be used as "labels" for training the classification models. [Clinical Dementia Ratings (CDR values: 0=non-demented; 0.5 = very mild dementia; 1 = mild dementia; 2 = moderate dementia)]
- Pandas is used for data loading and Python scikit-learn library for modeling.

The goal is to train machine learning models to predict whether the individuals in the cross-validation set (test set) have dementia ($CDR > 0$). The problem will be formulated both as a binary classification problem ($CDR = 0$, and $CDR > 0$). In the binary classification formulation, the $CDR > 0$ (i.e. $CDR = 0.5$, 1, and 2) values in the sliced dataset will be relabeled as $CDR = 1$,

The combined cross-sectional and longitudinal datasets after removal of rows with any column having an NaN value, has 339 records for $CDR = 0.0$, and 231 records for $CDR = 1$. Recall CDR values of 0.5 and 2.0 have been replaced by 1.0 to make it a binary classification problem. Thus the labels are "balanced", the number of records with $CDR = 0.0$ and $CDR = 1.0$ are 59% and 41% respectively.

- Classification Accuracy is used as the primary metric. This metric is applicable for binary classification where the number of records for the two labels is balanced. Classification accuracy is defined as the number of records correctly classified divided by the total number of records classified. Additionally, for binary classification here, AUC/ROC values are reported. Accuracy results are also be reported in sklearn Confusion Matrix format to evaluate classifier output quality, and in Classification Report format (provides, precision, recall, and f1-score) which are quite appropriate for the dataset used to train models for CDR classification. References 13 through 17 have details and discussion of these sklearn metrics.
- About 80% of the data in the dataset will be used for training the models. About 20% of data will used prediction of the CDR label for the k-fold cross-validation with $k=10$. Sensitivity studies with proportion other than 80:20, e.g. 70:30, will be considered to test sensitivity of this split on the accuracy.
- Data cleaning (e.g. removal of NaN values), data exploration, data preparation, data visualization, and data preprocessing will be described, as appropriate, and the impact of the latter on prediction metrics will discussed.

Metrics

Classification Accuracy is used as the primary metric. This metric is applicable for binary classification where the number of records for the two labels are balanced. As shown later in this notebook, the ratio of the number of processed records with CDR=0 and CDR=1 is 60% to 40%, and is considered balanced. Classification accuracy is defined as the number of records correctly classified divided by the total number of records classified. Additionally, for binary classification here, AUC/ROC values are reported. Accuracy results are also be reported in sklearn Confusion Matrix format to evaluate classifier output quality, and in Classification Report format (provides, precision, recall, and f1-score) which are quite appropriate for the dataset used to train models for CDR classification. References 13 through 17 have details and discussion of these sklearn metrics.

Benchmark

My primary benchmark is a neural network model (Appendix 1 in the Main Notebook) based on the same data discussed above and as used in this problem: Keras is used as the frontend with tensorflow backend. My secondary benchmark will be results of the study in the two papers below:

- Paper title: Usefulness of data from magnetic resonance imaging to improve prediction of dementia: population based cohort study, Reference 20.

"Results during 10 years of follow-up: there were 119 confirmed cases of dementia, 84 of which were Alzheimer's disease. The conventional risk model incorporated age, sex, education, cognition, physical function, lifestyle (smoking, alcohol use), health (cardiovascular disease, diabetes, systolic blood pressure), and the apolipoprotein genotype (C statistic for discrimination performance was 0.77, 95% confidence interval 0.71 to 0.82). No significant differences were observed in the discrimination performance of the conventional risk model compared with models incorporating data from MRI including white matter lesion volume (C statistic 0.77, 95% confidence interval 0.72 to 0.82; $P=0.48$ for difference of C statistics, Reference 21), brain volume (0.77, 0.72 to 0.82; $P=0.60$), hippocampal volume (0.79, 0.74 to 0.84; $P=0.07$), or all three variables combined (0.79, 0.75 to 0.84; $P=0.05$). Inclusion of hippocampal volume or all three MRI variables combined in the conventional model did, however, lead to significant

improvement in reclassification measured by using the integrated discrimination improvement index ($P=0.03$ and $P=0.04$) and showed increased net benefit in decision curve analysis. Similar results were observed when the outcome was restricted to Alzheimer's disease."

- Paper Title: The Use of MRI and PET for Clinical Diagnosis of Dementia and Investigation of Cognitive Impairment: A Consensus Report, Reference 22.

"Once the presence of dementia has been established, the role of imaging in the diagnosis of dementia subtypes is very much a function of the clinical diagnosis. The accuracy of the clinical diagnosis of Alzheimer's disease (AD) is quite good. Pathological AD has a prevalence of about 70% (range 50% to above 80% depending upon whether the AD occurs in isolation or with other entities) among all dementias (see evidence Table 1 in Reference 23); thus, even clinicians with limited neurological expertise should have a diagnostic accuracy, for AD at least, at about that level. A review of 13 published studies gave average values for sensitivity and specificity of the clinical diagnosis of AD of 81% and 70%, respectively(Reference 23). The overall accuracy of the clinical diagnosis of AD versus not-AD compared with the neuropathological standard based on those values for prevalence, sensitivity, and specificity, is 78%. "

Analysis

Dataset and Inputs

Reference 1 provides the downloadable MRI related data in comma separated value (csv) format. Reference 2 provides metadata and additional facts about the cross-sectional MRI.

OASIS Cross-sectional MRI Data in Young, Middle Aged, Non-demented and Demented Older Adults

- This dataset consists of a cross-sectional collection for 416 persons aged 18 to 96.
- For each person, 3 to 4 T1-weighted MRI scans that were obtained in single scan sessions are included.
- The persons include both men and women, and are all right-handed.
- In this dataset, one hundred persons over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD).
- Also, a reliability data set , Reference 3, is included which contains 20 non-demented subjects imaged on a subsequent visit within 90 days of their initial session.

- Dementia related ***Additional Data*** below for the cross-sectional MRI cases used this project. Features based on these ***Additional Data*** will be used to train classification models to predict the labels for the outcome (CDR).

OASIS: Longitudinal MRI Data in Non-demented and Demented Older Adults

This set consists of a longitudinal collection of 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included.

Note: MRI image pixel data are NOT used in this problem, only related features prefixed with @ sign (below) will be used.

The subjects are all right-handed and include both men and women.

72 of the subjects were characterized as non-demented throughout the study.

64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease. Another 14 subjects were characterized as non-demented at the time of their initial visit and were subsequently characterized as demented at a later visit.

Dementia related ***Additional Data*** below for the longitudinal MRI cases are used this project.

Features based on the ***Additional Data*** are relevant to finding machine learning solutions to the problem defined above, and will be used to train classification models to predict the labels for the outcome (Critical Dementia Rating, CDR).

Additional data: Specific References in parentheses below covering features are from Reference 2. These features include Demographic, clinical, and derived anatomic measures related to brain that are located in the file *oasis_crosssectional.csv*. Features prefixed with @ will be used for the problem.

Demographic data:

- @Gender (M/F), categorical data
- Handedness (Right or Left Handed), categorical data, all of which are right handed in the dataset.
- @Age (numeric),
- @Education (Educ, categorical). Education codes correspond to the following levels of education:

1=Less than high school graduate.

2=High school graduate.

3=Some college education

4=College graduate.

5=Beyond college.

Clinical data:

- **@Mini-Mental State Examination** (MMSE, Reference 8),

- **@Clinical Dementia Rating** (CDR, Reference 7)

0 = non-demented (**341** data points)

0.5 = very mild dementia (**193** data points)

-1 = mild dementia (**69** data points)

2 = moderate dementia (**5** data points)

There are some records with NaN values in one or more fields; these records will be removed from datasets prior to analysis. All participants with dementia (CDR >0) were diagnosed with probable Alzheimer's Disease. All participants with dementia (CDR >0) were diagnosed with probable Alzheimer's Disease.

Derived anatomic volumes data:

- **@Estimated total intracranial volume** (eTIV, mm3), Reference 4

- **@Atlas scaling factor** (ASF), Reference 4

- **@Normalized whole brain volume** (nWBV, mm3), Reference 5

Python and SK-Learn Libraries Loaded

Cross-sectional and longitudinal MRI related data were loaded from two CSV files. There are 436 rows and 12 columns of cross-sectional MRI related data and 373 rows and 15 columns of longitudinal MRI related data.

Load libraries

```
import numpy as np

from pandas import read_csv

from pandas.tools.plotting import scatter_matrix

from matplotlib import pyplot

from sklearn.model_selection import train_test_split

from sklearn.model_selection import KFold

from sklearn.model_selection import cross_val_score

from sklearn.metrics import classification_report

from sklearn.metrics import confusion_matrix

from sklearn.metrics import accuracy_score

from sklearn.linear_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

from sklearn.naive_bayes import GaussianNB

from sklearn.svm import SVC

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.ensemble import RandomForestClassifier
```

Data Exploration

Following type of information are obtained from data exploration. Additionally, graphical exploration of the datasets via plots provides more insights into data distribution, correlation among features (columns), and outliers (box plots).

- Dataset shape (number of rows and columns)
- Column headings
- Data Summary and statistics

- Data Types
- Count of data by classifier label (CDR) values

We note the following from the demographic and clinical features and cognitive test data for the cross-sectional MRI data, the following:

Mean Age for cross-sectional MRI data is about 51 years, mean for Education is 3.2 years, SES about 2.5, MMSE 27, eTIV 1481, nWBV 0.79, ASF 1.2.

Also note the missing (NaN) data from the count values. Those columns that have count of 436 have missing (NaN) data. The Delay column has most missing data followed by Educ, SES, MMSE, and the label (CDR) which have missing data in many rows. An option is to remove rows with missing column values, and that option will be chosen as seen later in this notebook. Another option, not used here, is to replace the NaN values with mean values listed for columns below. That latter choice would be appropriate if the column values have a normal or an uniform distribution. Also note that the values in the different columns are of different orders of magnitude, Some on the order of unity (SES, CDR, nWBV, and ASF), and other column values of higher of magnitude (Age and eTIV). Later in the notebook, rescaling or normalizing the columns (features) will be chosen as sensitivity studies for impact on accuracy and results.

Additional observations on the datasets:

- Note categorical data for gender M/F, and label (CDR) values ≥ 0.0 , and NaN. Cross-sectional MRI dataset has many rows with NaN values in multiple columns. These NaN values will be removed after merging the dataset with the Longitudinal MRI dataset.
- Data frequency for CDR label in the cross-sectional dataset: There are 135, 70, 28, 2, and 201 rows with CDR label values 0, 0.5, 1.0, and 2, and NaN. We see only 2 data rows for CDR=2.
- The three additional columns (Subject ID, Group, and Visit) in longitudinal dataset are not in the cross-sectional dataset. These three columns are not meaningful for the CDR classification, and will be dropped before merging the two datasets. The MR Delay and Delay columns in the two datasets have same meaning and the MR Delay column will be repositioned to be in the same column order as the Delay column.
- From the descriptive statistics of the longitudinal dataset we find the following: Mean Age for longitudinal MRI data is about 77 years, mean for Education is 14.6 years, SES about 2.5, MMSE 27, eTIV 1488, nWBV 0.73, ASF 1.2. Also note the missing (NaN) data from the count values. Those columns that have count of 436 have missing (NaN) data. The Delay column has most missing data followed by Educ, SES, MMSE, and the label (CDR) which have missing data in many rows. Another option, not used here, is to replace the NaN values with mean values listed for columns below. That latter choice would be appropriate if the column values have a normal or an uniform distribution. Also note that the values in the different columns

are of different orders of magnitude, some on the order of unity (SES, CDR, nWBV, and ASF), and other column values of higher of magnitude (Age and eTIV).

- Longitudinal MRI dataset has many rows with NaN values in multiple columns. These NaN values will be removed after merging the dataset with the Cross-sectional MRI dataset.
- For the longitudinal MRI dataset, there are no records with missing values (NaN) in the CDR column. Earlier we found 201 rows with missing values in the cross-sectional dataset.

Data frequency for CDR label:

- In the longitudinal MRI dataset, there are 206, 123, 41, and 3 rows with CDR label values 0, 0.5, 1.0, and 2, respectively.
- There are very few (3) data points with CDR=2. This along with no CDR=2 data in the cross-sectional dataset, is a potential limitation in modeling the classification problem as multi-label (not binary) classification since the some cross validation sets may have very few or even no data points. Metrics for multi-label classification may not be meaningful for CDR=2 data. One alternative is to not include the few CDR=2 data points in the multi-label classification case. Another alternative is to duplicate the existing CDR=2 data rows a few times and append to the combined dataset.
- The count of various CDR labels for the two datasets and the total values are shown in the Table below. Again we see that the total number of records for the combined dataset would be 5, a small number of data records to calculate meaningful metrics.

CDR →	0	0.5	1	2	NaN
Cross-Sectional Dataset	135	70	28	2	201
Longitudinal Dataset	206	123	4	3	0
Total	341	193	69	5	201

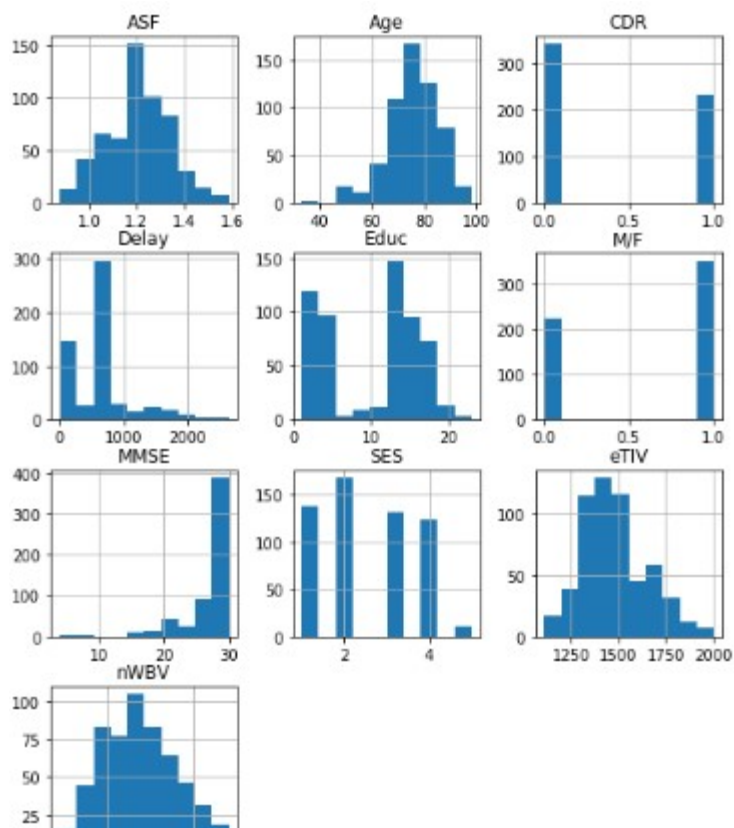
Data Preprocessing

- Replace gender categorical data (M, F) with numerical values (0, 1).
- Select longitudinal dataset (dfoasl) columns that have data similar to the dataset for cross-sectional dataset (dfoasx). This will facilitate combining the two datasets to create a merged dataset (dfoas_merge) with 809 rows and 12 columns.
- Explore the merged dataset
- Remove NaN rows; drop rows with NaN values (missing data) in at least one column. This will make classification metrics meaningful.

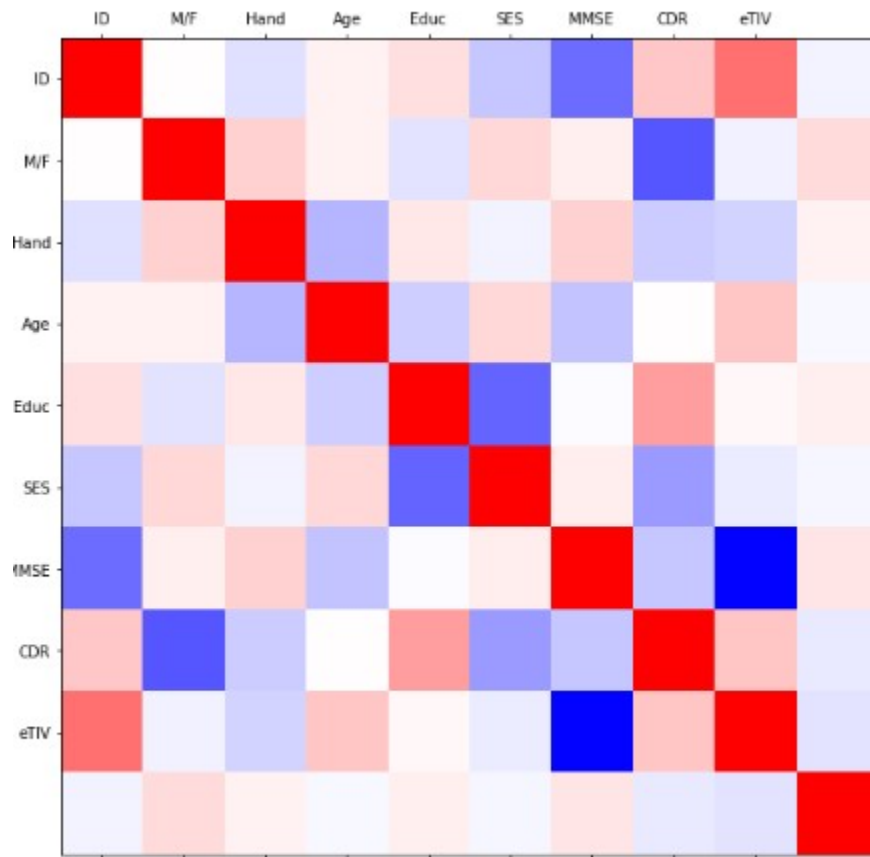
- Next, we drop rows in the merged dataset which have NaN in any column. We see that there are $809-570=239$ rows with NaN values (missing data) in at least one column.
- The merged dataset has 570 rows and 12 columns.
- Convert CDR>0 values to 1 to make this a binary classification problem (i.e. CDR values: 0 or 1). That is, leave CDR values of 0 as is and convert CDR values >0 to 1.
- **Binary Labels**(CDR values: 0,1): We see from above exploratory results that the combined cross-sectional and longitudinal datasets have 339 records for CDR=0.0, and 231 records for CDR=1 (recall CDR values of 0.5 and 2.0 have been replaced by 1.0 to make it a binary classification problem. Thus the labels are "balanced", the number of records with CDR=0.0 and CDR=1.0 are 59% and 41% respectively. *This justifies the use of Classification accuracy as a metric in addition to ROC/AUC, Confusion Matrix (sensitivity, specificity, recall, and support metrics) from sklearn.*

Exploratory Visualization

- The following visualizations of the merged dataset (dfoas_merge dataframe) have been included in the Main Jupyter Notebook. The histogram is shown below..
 - ✓ Histogram
 - ✓ Density Plots
 - ✓ Box Plots
 - ✓ Scatter Plots
 - ✓ Correlation matrix
- From the histogram, density plots, and box plots for the feature variables, we note the following: Age, eTIV, nWBV, and ASF have approximately normal distribution. The feature variables, Educ, and SES have bi-modal or multimodal distribution. As expected CDR after transformation of CDR>0 values to 1, indicates binary distribution (0, and 1). The gender variable M/F indicates transformed numerical values of 0 and 1 as expected.



- From the Correlation plot we see that We see that SES and Educ are correlated and MMSE and eTIV are correlated



Benchmark

My primary benchmark will be a neural network model based on the same data discussed above and as used in this problem: Will use Keras frontend and tensorflow backend. My secondary benchmark will be results of the study in the two papers below:

1. Paper title: Usefulness of data from magnetic resonance imaging to improve prediction of dementia: population based cohort study <http://www.bmj.com/content/350/bmj.h2863>

"Results During 10 years of follow-up, there were 119 confirmed cases of dementia, 84 of which were Alzheimer's disease. The conventional risk model incorporated age, sex, education, cognition, physical function, lifestyle (smoking, alcohol use), health (cardiovascular disease, diabetes, systolic blood pressure), and the apolipoprotein genotype (C statistic for discrimination performance was 0.77, 95% confidence interval 0.71 to 0.82). See Reference 21 for definition of

c-statistics. No significant differences were observed in the discrimination performance of the conventional risk model compared with models incorporating data from MRI including white matter lesion volume (C statistic 0.77, 95% confidence interval 0.72 to 0.82; $P=0.48$ for difference of C statistics), brain volume (0.77, 0.72 to 0.82; $P=0.60$), hippocampal volume (0.79, 0.74 to 0.84; $P=0.07$), or all three variables combined (0.79, 0.75 to 0.84; $P=0.05$). Inclusion of hippocampal volume or all three MRI variables combined in the conventional model did, however, lead to significant improvement in reclassification measured by using the integrated discrimination improvement index ($P=0.03$ and $P=0.04$) and showed increased net benefit in decision curve analysis. Similar results were observed when the outcome was restricted to Alzheimer's disease."

2. Paper Title: The Use of MRI and PET for Clinical Diagnosis of Dementia and Investigation of Cognitive Impairment: A Consensus Report

https://www.alz.org/national/documents/imaging_consensus_report.pdf

"Once the presence of dementia has been established, the role of imaging in the diagnosis of dementia subtypes is very much a function of the clinical diagnosis. The accuracy of the clinical diagnosis of Alzheimer's disease (AD) is quite good. Pathological AD has a prevalence of about 70% (range 50% to above 80% depending upon whether the AD occurs in isolation or with other entities) among all dementias (see evidence Table 1 in reference 1); thus, even clinicians with limited neurological expertise should have a diagnostic accuracy, for AD at least, at about that level. A review of 13 published studies gave average values for sensitivity and specificity of the clinical diagnosis of AD of 81% and 70%, respectively (1). The overall accuracy of the clinical diagnosis of AD versus not-AD compared with the neuropathological standard based on those values for prevalence, sensitivity, and specificity, is 78%. "

Methodology

Solution Statement

- Train a supervised machine learning classification model to properly classify the OASIS data according to clinical dementia ratings(CDR values).
- Train a number of candidate models from the scikit-learn library (Reference 12) such as Logistic Regression, Linear Discriminant Analysis, KNN, Naive Bayes, CART, and SVM.
- Select the best model based on Classification "Accuracy", as metric. Additionally, for this binary classification AUC/ROC values will be reported. For multi- class classification (multiple

CDR labels) F-1 score will be reported. The results from the best model will be provided along with those from the other models.

- Combine the cross-sectional and longitudinal MRI related demographic and clinical data into a single dataset. This increases number of data points for training the classifier
- Split the resulting dataset into training dataset (80%) and the remaining data(20%) for typical ten-fold cross validation.
- Report the prediction accuracy of the models and identify the model that yields the highest classification accuracy. Report accuracy results in sklearn Confusion Matrix format (to evaluate classifier output quality) and, Classification Report format (provides, precision, recall, f1-score). See References 13 through 17.

Features Chosen for Training

The following seven features (merged dataset columns) were chosen for training the classifiers to and to predict the labels (CDR variable) in the validation set. The column indices of these features are shown to the left of the features. A 80:20 split was used for the training vs. validation sets. When the split was changed to 70:30, the results (metrics) only slightly.

Column Index	Feature Variable	Label Variable
1	M/F	
3	Age	
4	Educ	
5	SES	
6	MMSE	
7		CDR
8	eTIV	
9	nWBV	
10	ASF	

Algorithms/Classifiers

A combination of eight linear and non-linear algorithms (classifiers) were used to spot check (evaluate) validation accuracy. These are:

1. Logistic Regression (LR).
2. Linear Discriminant Analysis (LDA).
3. K-Nearest Neighbors (KNN).
4. Classification and Regression Trees (CART).
5. Gaussian Naive Bayes (NB).
6. Support Vector Machines (SVM).
7. Gradient Boosting Machine (GBM).
8. Random Forest Classifier (RFC).

This list is a good mixture of simple linear (LR, LDA), and nonlinear (KNN, CART, NB and SVM) algorithms. We reset the random number seed before each run to ensure that the evaluation of each algorithm is performed using exactly the same data splits. It ensures the results are directly comparable.

The following results were obtained for the mean and standard deviation of cross validation accuracy of the eight classifiers. The Main Jupyter notebook has the details. Two of the classifiers, Random Forest (RFC) and Gradient Boosting (GBM) yielded the highest accuracies (~0.86). These two classifiers were further used with additional details of the algorithms to obtain detail metrics in addition to accuracy, e.g. Precision, Recall, F-1 Score and AUC (Area under the curve) values. The coding and parameter selection for these cases are documented in detail in the Main Jupyter notebook for binary classification (CDR labels 0,1) and in the Multilabel Jupyter notebook for the multi-label CDR validation cases. The results are summarized next

Model	Spot-Check Results → Cross-Validation Accuracy: (Mean, Stdev)
LR	(0.80, 0.05)
LDA	(0.80, 0.04)
KNN	(0.64, 0.04)
CART	(0.80, 0.07)
NB	(0.81, 0.06)
SVM	(0.62, 0.07)
GBM	(0.86, 0.05)
RFC	(0.86, 0.06)

Results

Classifier	Labels: CDR values in dataset	Accuracy	Precision (avg)	Recall (avg)	F1- Score (avg)	AUC	Comment
Benchmark: Neural Network: 4-node input layer, two 8-node hidden layers and one 1- node Output layer	(0,1)	0.78	N/A	N/A	N/A	N/A	See Jupyter Notebook with the word “Benchmark” in the name, which was executed on Floydhub GPU
Random Forest classifier	(0,1)	0.886	0.89	0.89	0.88	0.901	Binary Classification for CDR
Gradient Boosting Classifier	(0,1)	0.895	0.89	0.89	0.89	0.915	Binary Classification for CDR
PCA 3 components, select 4 best, Decision Tree	(0,1)	0.807	N/A	N/A	N/A	N/A	Binary Classification for CDR
PCA 3 components, select 4 best, Gradient Boosting	(0,1)	0.840	N/A	N/A	N/A	N/A	Binary Classification for CDR
Gradient Boosting Classifier	(0, 0.5, 1, 2)	0.81	0.82 0.81 0.89 0.50	0.95 0.59 0.67 1.0	0.88 0.69 0.76 0.67	N/A	Multi-label Classification for CDR; appended rows with CDR=2 to the dataset to artificially increase CDR=2 occurrence to 15 rows from 5 rows.

Free-Form Visualization (Learning Curves)

Learning curves are provided in **Appendix 2** in the Jupyter Notebook (name ending in _MAIN). These curves provide visual demonstration of the dependence of the training and validation accuracy on the size of the dataset (number of records or samples) for the merged OASIS dataset for a number of classifiers such as Random Forest and Gradient Boosting. This visualization clarifies the number of samples required for a particular prediction accuracy threshold, and qualitatively/visually demonstrates the level of overfitting in the training dataset.

Conclusions, Justification, and Reflections

- The formulation of OASIS data (Ref 1 and 2) in terms of a dementia classification problem based on demographic and clinical data only (and without directly using the MRI image data), is a simplification that has major advantages and appeal. This means the trained model can classify whether an individual has dementia or not with about 89% accuracy, without having to wait for radiological interpretation of MRI scans. This can provide an early alert for intervention and initiation of treatment for those with onset of dementia.
- The assumption that the combined cross-sectional and longitudinal datasets would lead to dementia label (CDR) classification with acceptable accuracy turned out to be true. The method required careful data cleaning and data preparation work, converting it to a binary classification problem, as outlined in this notebooks.
- At the outset it was not clear which algorithm(s) would be more appropriate for the binary and multi-label classification problem. The approach of spot checking the algorithms early for accuracy led to the determination of a smaller set of algorithms with higher accuracy (e.g. Gradient Boosting and Random Forest) for a deeper dive examination, e.g. use of a k-fold cross-validation approach for classifying the data correctly.
- The neural network benchmark model accuracy of 78% for binary classification was exceeded by the classification accuracy of the main output of this study, the trained Gradient Boosting and Random Forest classification models (89%). This builds confidence in using the latter model to classify new data, e.g. data from new patients.

References

1. The Open Access Series of Imaging Studies (OASIS), <http://www.oasis-brains.org/app/template/Index.vm;jsessionid=6926BBF18A3D5CD974E750FAC8ED01CE>
2. OASIS Fact Sheet (rev. 2007-8-20) Cross-Sectional Data Across the Adult Lifespan, Marcus et al., 2007, http://www.oasis-brains.org/pdf/oasis_cross-sectional_facts.pdf
3. MRI Reliability data across the adult lifespan, http://www.oasis-brains.org/app/action/BundleAction/bundle/OAS1_RELIABILITY
4. Buckner, RL, Head, D, Parker, J, Fotenos, AF, Marcus, D, Morris, JC, Snyder, AZ, 2004. A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. Neuroimage 23, 724-38.

5. Fotenos, AF, Snyder, AZ, Girton, LE, Morris, JC, and Buckner, RL, 2005 Normative estimates of crosssectional and longitudinal brain volume decline in aging and AD. *Neurology*, 64: 1032-1039.
6. Marcus, DS, Wang, TH, Parker, J, M, Csernansky, JG, Morris, JC, Buckner, RL, 2007. Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults. *Journal of Cognitive Neuroscience*, 19, 1498-1507.
7. Morris, JC, 1993. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* 43, 2412b-2414b.
8. Rubin, EH, Storandt, M, Miller, JP, Kinscherf, DA, Grant, EA, Morris, JC, Berg, L, 1998. A prospective study of cognitive function and onset of dementia in cognitively healthy elders. *Arch Neurol*. 55, 395- 401.
9. Zhang, Y, Brady, M, Smith, S, 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. *IEEE Trans. on Medical Imaging*, 20(1):45-57.
10. Diagnosis of Alzheimer's Disease Based on Structural MRI Images Using a Regularized Extreme Learning Machine and PCA Feature
<https://www.hindawi.com/journals/jhe/2017/5485080/>
11. Use of Machine Learning Technology in the Diagnosis of Alzheimer's Disease
[http://doras.dcu.ie/21356/1/Noel_s_Masters_thesis_Copy_\(1\).pdf](http://doras.dcu.ie/21356/1/Noel_s_Masters_thesis_Copy_(1).pdf)
12. Scikit-learn, <http://scikit-learn.org/stable/index.html>
13. Model evaluation: quantifying the quality of predictions, http://scikit-learn.org/stable/modules/model_evaluation.html.
14. Precision and Recall, http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html#sphx-glr-auto-examples-model-selection-plot-precision-recall-py
15. Confusion Matrix, http://scikit-learn.org/stable/modules/model_evaluation.html#confusion-matrix
16. Support, http://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html
17. F1-score, http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html
18. Conditional data slicing in a Pandas dataframe
<https://stackoverflow.com/questions/17071871/select-rows-from-a-dataframe-based-on-values-in-a-column-in-pandas>

19. Matplotlib colormap examples and color schemes for using in heatmap:
<http://pyhogs.github.io/colormap-examples.html>;
https://matplotlib.org/examples/color/colormaps_reference.html
20. Usefulness of data from magnetic resonance imaging to improve prediction of dementia: population based cohort study <http://www.bmj.com/content/350/bmj.h2863>
21. C - Statistics: <http://www.statisticshowto.com/c-statistic/>
22. The Use of MRI and PET for Clinical Diagnosis of Dementia and Investigation of Cognitive Impairment: A Consensus Report
https://www.alz.org/national/documents/imaging_consensus_report.pdf
23. Knopman DS, DeKosky ST, Cummings JL, Chui H, Corey-Bloom J, Relkin N, et al. Practice parameter: Diagnosis of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology 2001;56(9):1143–1153.
24. Machine Learning Mastery, <https://machinelearningmastery.com/>
25. "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics", <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3706434/>
26. "8 Proven Ways for improving the “Accuracy” of a Machine Learning Model",
<https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/>
27. “A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning”,
<https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/>
28. "Gradient Boosting Tree vs. Random Forest",
<https://stats.stackexchange.com/questions/173390/gradient-boosting-tree-vs-random-forest>
29. "Does the dataset size influence a machine learning algorithm?", <https://stackoverflow.com/questions/25665017/does-the-dataset-size-influence-a-machine-learning-algorithm?rq=1>
30. http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html#sklearn.feature_selection.SelectKBest
31. "Plotting Learning Curves", http://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html#sphx-glr-auto-examples-model-selection-plot-learning-curve-py

Appendix 1- Files

The following files related to the Capstone Analysis, have been submitted to Udacity for review:

#	File Name	File Type	Note	Note
1	Review_Final_MLND Capstone Project PROPOSAL.pdf	PDF	Udacity Review of Capstone Proposal	“Meets Specification”
2	MLND Capstone Analysis Report Alamgir.docx/pdf/html	Word/PDF/html	Capstone Project Analysis Report	This document
3	MLND Capstone Project Analysis - Rubric-OASIS MAIN Alamgir.ipynb/pdf	Jupyter Notebook	Binary Label (CDR=0, 1) Classification Analysis file.	Look at this File first!
4	MLND Capstone Project Analysis - Rubric-OASIS MULTILABEL Alamgir.ipynb	Jupyter Notebook	Multi-label (CDR=0, 0.5, 1.5, 2) Classification Analysis file.	This “MULTILABEL” file models multiple CDR labels vs. two labels in File “MAIN”
5	MLND Capstone Project Analysis - Rubric-OASIS BENCHMARK_ANN_4-8-8-1_Keras_Py27_Tensorflow .ipynb	Jupyter Notebook/PDF	Contains Benchmark analysis using neural network 4-8-8-1 model. Uses GPU processing on floydhub (www.floydhub.com) with KERAS front end and Tensorflow backend running on Python 2.7	4-8-8-1 indicates a neural network model with an input layer with 4 nodes, two hidden layers with 8 nodes each, and one output layer. RELU activation function used for the hidden layers and Sigmoid for the output layer.
6	oasis_cross-sectional.csv	CSV	CSV file containing OASIS cross-sectional MRI related data used in the Jupyter notebooks above.	Contains clinical and demographic data for patients, no image data
7	oasis_longitudinal.csv	CSV	CSV file containing OASIS longitudinal MRI related data used in the Jupyter notebooks above.	Contains clinical and demographic data for patients, no image data

