Predicting Dementia

Reduce monetary cost of dementia with early diagnosis by predicting dementia

Ozkan Serttas - February 7, 2019

Introduction

In this project, we will work with two different datasets. First one is collected from CDC called Healthy Aging Data which is a survey data and subjects give information about their health condition. We use this data to see impact (if exists) of Dementia problem in the U.S. If not the study will stop there so no further analysis would be needed.

Image was taken from CDC

If the answer is yes, then a further analysis on dementia is needed to understand its factors. To explore factors of dementia we will use data from longitudinal study on determining cognitive functionality performance of subject. With the second data in hand, the final objective is to build a model that can classify cognitive impairment (being demented) successfully and then interpret the results to find factors which influence the score.

Motivation

Dementia is a general term for a decline in mental ability severe enough to interfere with daily life. Memory loss is an example.
Dementia is not a specific disease. It's an overall term that describes a group of symptoms associated with a decline in memory or other thinking skills severe enough to reduce a person's ability to perform everyday activities . As Alzheimer.org repots that every seconds someone in the U.S. develops Alzheimer's disease. Moreover, in the same report we see that from 2000 to 2015 there was an increase of in deaths due to Alzheimer's disease which can be called later stages of cognitive impairment.
Although not every cognitive impaired patients progress to Alzheimer's, cognitive impairments called as Early Alzheimer's especially the severe ones. Study ***Alzheimer's org*** shows that the Medicare spending per person is about $7415 for seniors without Alzheimer's and other types of dementia whereas it is about $24,122 per person in seniors with other diseases. Reports also indicate that the Medicaid money spent on dementias like Alzheimer six times higher than other money spent on medical problems for 65 year old or above patients. Moreover, Alzeimer.org reports indicate that from 2000 to 2015 there was an increase of 123% in deaths due to Alzheimer's disease which can be called later stages of cognitive impairment.

Approach

Data Wrangling phase: Import and inspect raw data. Isolate relevant variables, fill or calculate new variables, and organize the dataframe. Resolve missing, invalid, corrupted, duplicate values.

Exploratory Data Analysis: Explore datasets using statistical analysis methods to explain variables. Use help of visualization methods to identify statistical features of datasets.

Modeling: As for the problem structure, purpose of modeling is to classify subjects whether or not subject is demented based on features in the dataset. In statistics, this type of problems go under Classification problems where algorithm tries to find best decision boundary for a given probability threshold. Default probability threshold of 0.5 is used herein modeling section to identify if subject is demented or not. After setting permanence metrics as DOR

score and Accuracy, to identify best performing model hyper-parameter tuning and other model selection techniques are applied.

Data Wrangling

The Datasets

First dataset, Healthy Aging Data is collected from CDC website in csv format.This data collected by The_Behavioral_Risk_Factor_Surveillance_SystemBRFSS which contains results of health related **telephone surveys** to monitor health-related risk behaviors, chronic health conditions, and use of preventive services of the US residents
The second dataset is downloaded from Kaggle website which was in csv format and can be reached from this link: https://www.kaggle.com/jboysen/mri-and-alzheimers
Second dataset consists of a longitudinal collection of 150 subjects aged 60 to 96. In this longitudinal study each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions.

- Healthy Aging dataset from CDC:
  - 29 Columns and 20,000 rows in the original data
  - Covers surveys from roughly 50 U.S. states for years 2011, 2015 and 2016
  - People from four different age groups were surveyed and asked questions about their health conditions

- Second dataset MRI scans data from OASIS-Brains :
  - 150 Subjects aged between 60 to 96
  - 373 rows and 15 columns
  - For each subject 3 or 4 individual T1-weighted MRI scans obtained in a single session(visit)
  - Data includes demographic information about subjects such that their socioeconomics, education levels, genders etc.

Variables of OASIS data

- Group :Demented/Un-demented/Converted)
- Visit :Number of visits
- eTIV - Estimated total intracranial volume, mm3
- nWBV - Normalized whole-brain volume, expressed as a percent of all voxels in the atlas-masked image that are labeled as gray or white matter by the automated tissue segmentation process
- ASF - Atlas scaling factor (unit-less). Computed scaling factor that transforms native-space brain and skull to the atlas target (i.e., the determinant of the transform matrix)
- Gender: (M/F)
- Educ : Years of education
- Age : 60 to 96
- SES : Socioeconomic status as assessed by the Hollingshead Index of Social Position and classified into categories from 1 (lowest) to 5 (highest)
- MMSE : Mini mental stage examination has range from 0 = worst to 30 = best
- Clinical Dementia Rating has range 0 to 3. Above 0.5 is condisered probable AD
- eTIV/(ICV) : Estimated total intracranial volume
- ASF : Atlas scaling factor

$$eTIV = \frac{ICV_{Atlas}}{ASF}$$

Data Cleaning on CDC Data

Cognitive Impairment Data extracted from CDC Healthy Aging Survey Data to only get subjects with cognitive issues. After extraction we have left with 29 columns and 3696 observation. To clean missing data we set 50 % threshold and drop a column if more than 50 % of a column is missing. Remaining missing values are imputed with using appropriate statistic value based on the data-type.

Data Cleaning on OASIS Data

As the dataset obtained from a research facility it is almost tidy. Missing values are filled with single imputation technique which is using appropriate statistic (mean, median or mode). Redundant columns meaning columns that have zero variance are removed SES column is reversed to convey the right information to the model. It was originally defined as low value showing high economical status and high value for low income status which might mislead the model. To prevent that the order is reversed.

CDR column is removed before stepping into modeling to prevent Data-Leakage, as Diagnosis Group column derived form CDR column. Categorical columns if exist encoded as numeric values. Order of Ordinal Data is fixed to prevent misinterpretations. Usually low level should be coded with smaller integer than higher level category.