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

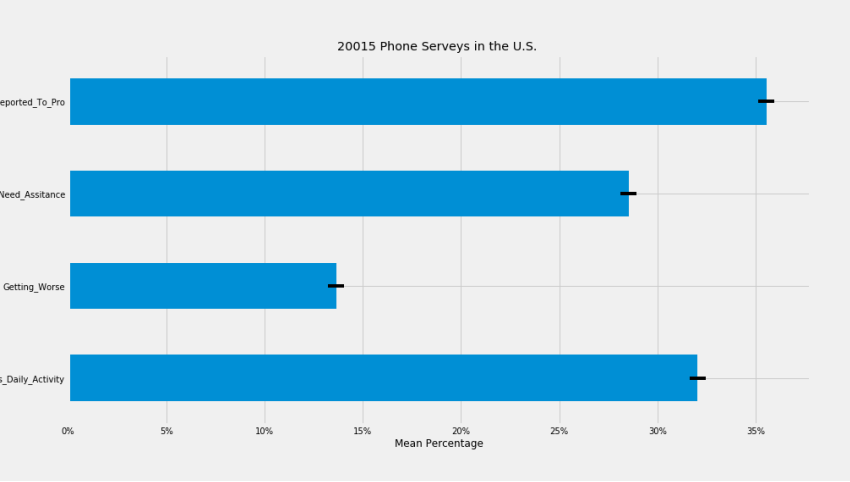
As the dataset obtained from a research facility it is almost tidy. Missing values are filled with single imputation technique which is using appropriate statistic (mean, median or mode). Redundant columns meaning columns that have zero variance are removed SES column is reversed to convey the right 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.

Exploratory Data Analysis

Explore datasets using statistical analysis methods to explain variables. Use help of visualization methods to identify statistical features of datasets. A major part of EDA is searching for relationships between the features and the target. Investigating covariant variables especially ones have significant positive or negative correlation between target_variable is an important task.

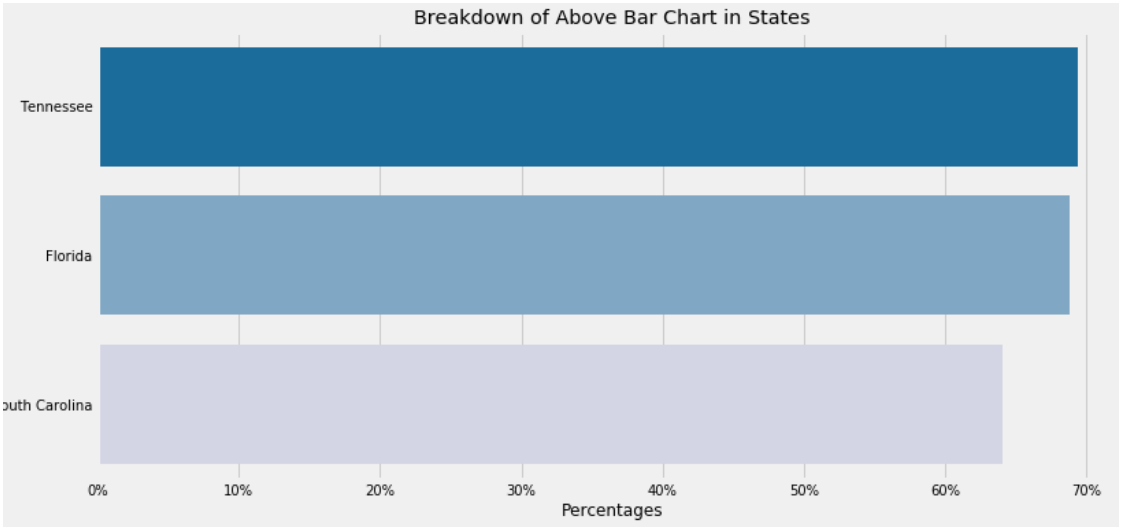
CDC Dataset



Average percentages show that Cognitive health issues are affecting patients' life condition severely. About 30 % of the patients need assistance to take care of daily activities while 32 % reports considerable effects of SCD.
Subjects who talked to a professional about their condition makes up 35 % overall states in the U.S.

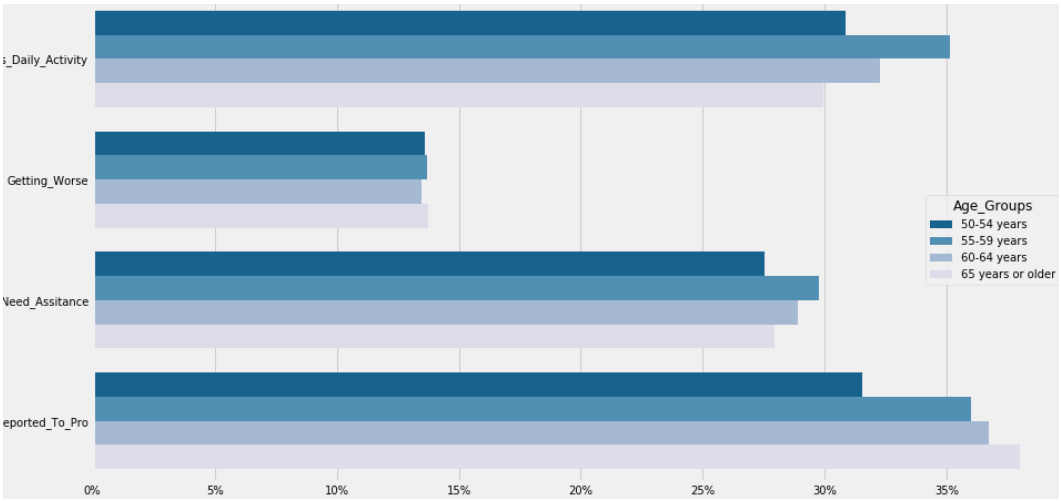
Subjective Cognitive Impairment Survey Data

Is there a decline in cognitive decline related diseases? or in the number of reports? As Alzheimer.org repots that every 65 seconds someone in the U.S. develops Alzheimer's disease. Moreover, from the same report we see 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



Top three states that have most subjective cognitive complaints

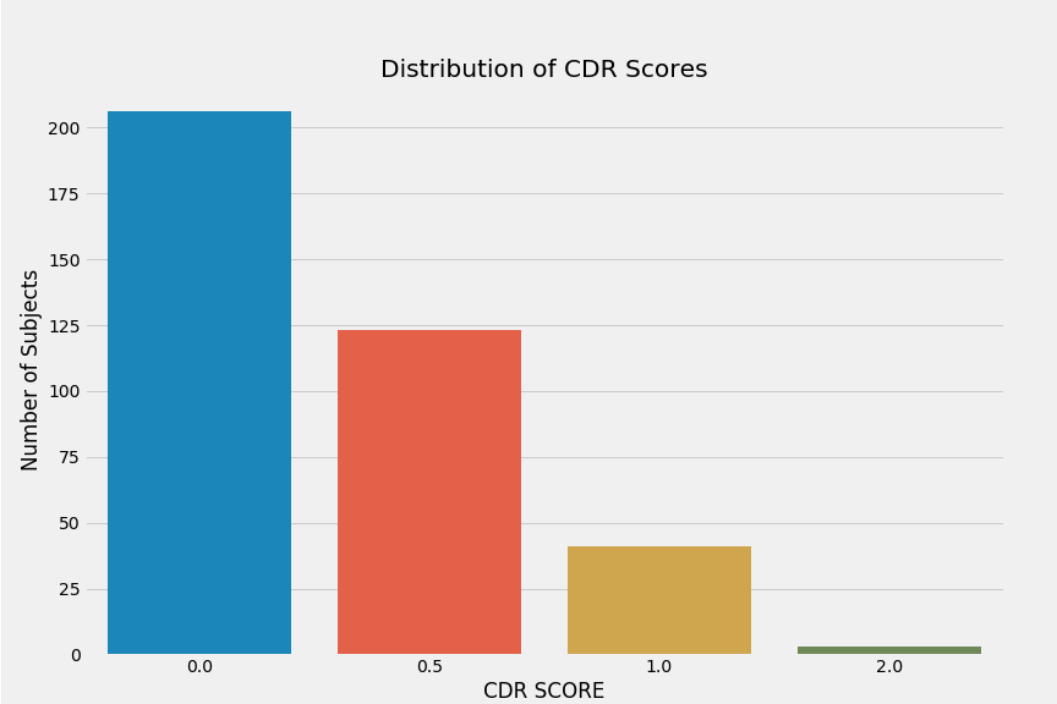
impairment. Although not every cognitive impaired patients progress to Alzheimer's, cognitive impairments called as Early Alzheimer's especially the severe ones.



Breakdown of Above Chart into Age Groups

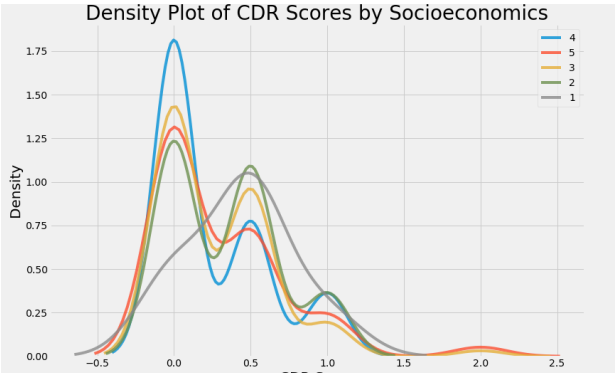
OASIS-Brains Dataset

One way to examine the effect of a categorical variable on the target is through a density plot. A density plot can be thought of as a smoothed histogram because it shows the distribution



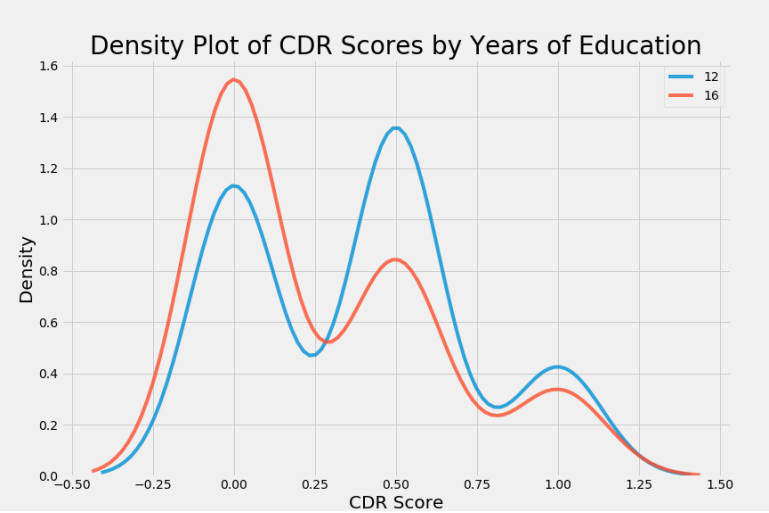
0 is showing no signs of dementia, 0.5 is for probable Alzheimer's Disease and so on.

of a single variable. We can color a density plot by class to see how a categorical variable changes the distribution. The CDR score, a dementia staging tool, is used to indicate the level of dementia. All subjects with CDR > 0 are diagnosed with probable AD(Alzheimer's Disease.



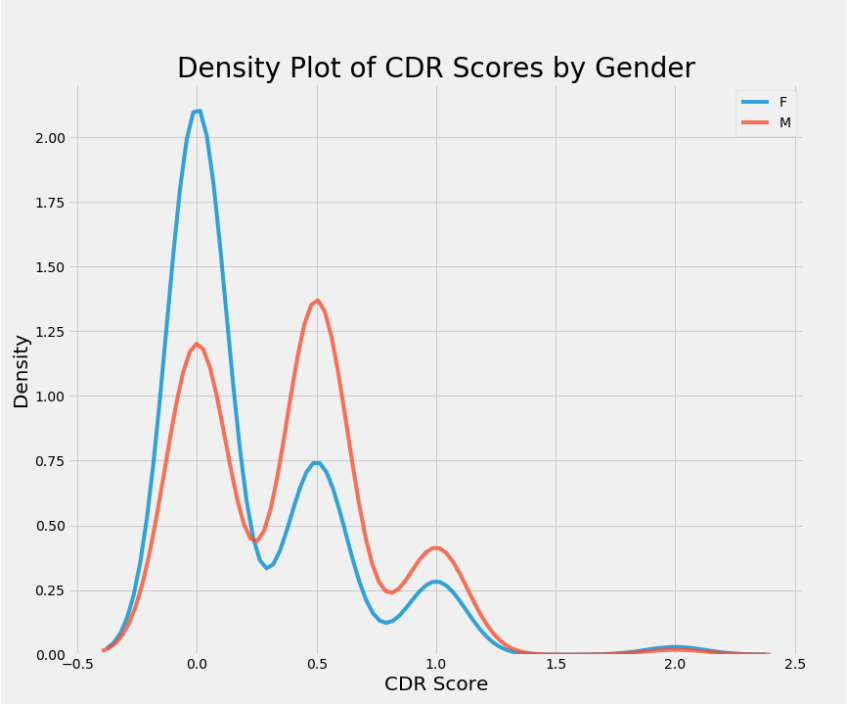
We can see that CDR score of 0 are more common in low-SES group (1) whereas for high-SES group (5) we see that CDR score of 0.5 is more common than others.

Affect of Socioeconomics Level of Subject on CDR



Distribution of Years of Education by CDR

Subjects with 12 years of education slightly have high CDR score than the ones with 16 years of education.



Distribution of Gender by CDR

Majority of females seems to be healthier than males. Males seem to distributed almost every phase of the CDR scores. We may think of using **Stratified Sampling** to investigate gender affect more when we are splitting our dataset before applying ML algorithms.

Hypothesis Testing

Null Hypothesis is being there is no difference in respect to Gender. Females and Males are from the same distribution.

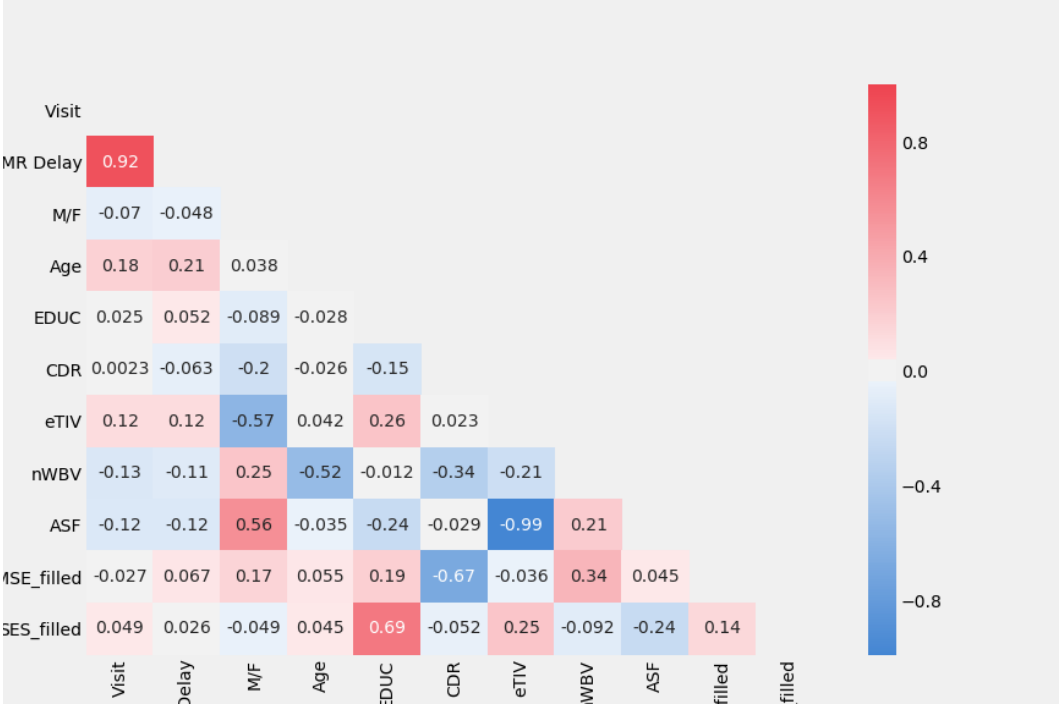
T-Test Results:

p_value = 3.376e-7

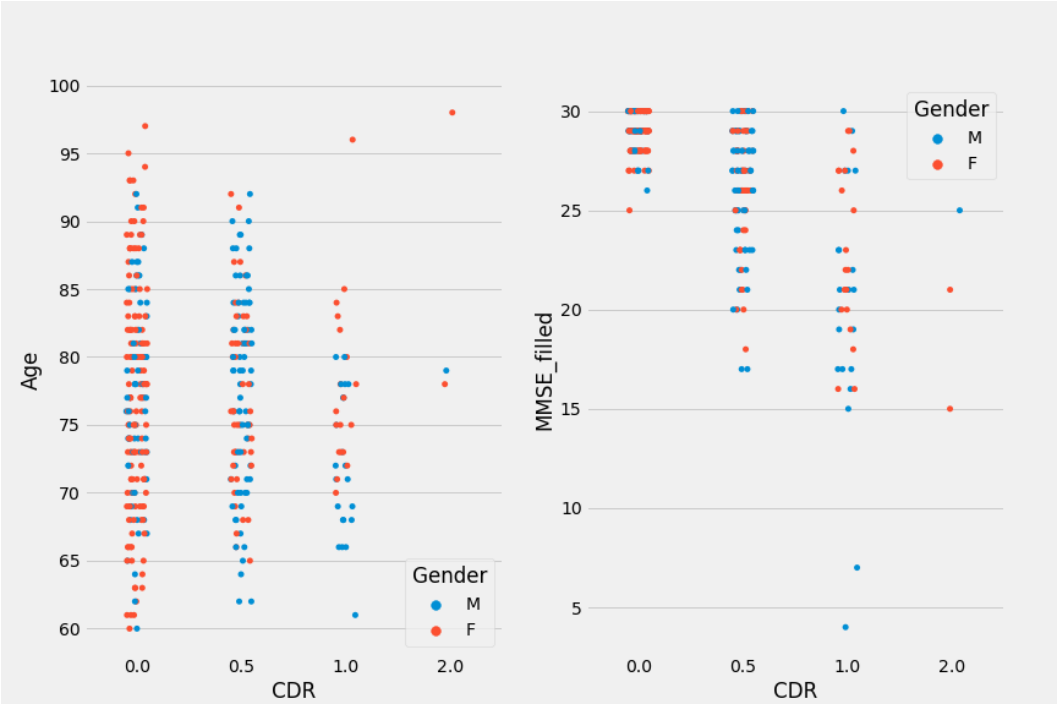
t_statistic = -5.174

Student t test result suggests that the samples are from different distributions. There is significant difference between genders of

subjects being diagnosed Demented. This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances by default.



Discovering Linear Relations If Exists



Age and MMSE vs CDR by Gender

Between ASF and eTIV a highly negative correlation exists while we observe highly positive correlation between Visit and MR Delay .
Briefly, there is not a pattern that suggests change in Age influences the change in CDR score. The second plot suggests that the MMSE scores below 25 result in dementia, but the scores above 25 is spread over demented and non-demented classes so the scores above 25 are not distinctive.

Behaviors of Continuous Variables by Gender

Recap of Preliminary Analysis of Features

`MMSE`: Based on EDA process over all , we may infer that a high `MMSE` scores covariate with a small `CDR` scores. This indicates that `MMSE` has influence on classification of the subject.

`SES`: If we look at the `SES` socioeconomic statue column, we can observe a high variation in its classes and also low (.05) correlation coefficient supports this phenomena. So we may suggest that `SES` class does not look like have an influence on the `CDR` score.

`GENDER`: Looks like gender is an influential factor on CDR score or for a subject to get diagnosed with dementia.

`ASF eTIV nWBV`: While `ASF` Atlas scaling factor and `eTIV` estimated total intracranial volume have negative correlation between them, other pairs seem to uncorrelated. The negative correlation exists because of the computational process of eTIV which is computed as $eTIV = \frac{ICV_{Atlas}}{ASF}$ according to FreeSurfer website.

`Visit`: We have observed high positive correlation between `Visit` number of visits and `MR_Delay` so we can drop one of them as they are basically providing the similar information to the model. In other words we need to get collinearity out of the way.

`EDUC SES Gender` has some effect on CDR score but there is not any strong relationship.

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.

Summary of Steps in Modeling

- Created a binary Diagnosis column based on CDR scores for each observation which to be used as binary class.

- Dataset is split into **Train** and **Test** split using **StratifiedKFold** algorithm of **Sklearn** library. It is to make sure the Test set have samples from all classes present in Training set.

- To measure models performance, DOR **Diagnostic Odds Ratio metric** is adopted. DOR measures the effectiveness of a diagnostic test. It is defined as the ratio of the odds of the test being positive if the subject has a disease relative to the odds of the test being positive if the subject does not have the disease.

- Three types of algorithms are used to determine the best performing classifier:

1. Linear Models
   - Logistic Regression
   - Linear Discriminant Analysis

2. Non-Linear Models
   - Neural Networks
   - SVM with Linear Kernel
   - Naive Bayes

3. Tree Based Models
   - Basic Decision Tree
   - Random Forest Classifier
   - Gradient Boosting
   - XGBoost Classifier

To compare models, two performance metric were used even though more than two were shown. The first metric is used DOR Diagnostic Odds Ratio score which measures effectiveness of a diagnostic test. Second metric is used Accuracy score on test set. We first choose top 3 models with the highest DOR score than compare Accuracy Scores among those selected ones.