CS109 Final Project - Milestone 3 - EDA report

Shristi Pandey, James May, Zachary Werkhoven, Guilherme Braz

Data description

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a study that aims to understand factors leading to the Alzheimer's Disease (AD) and to track disease progression. The study began in 2004 and has progressed through 4 phases(ADNI1, ADNI-GO, ADNI-2, and ADNI-3) with distinct research goals. The central focus of ADNI is to provide neuroanatomical imaging on patients over time, but the study also includes comprehensive information from genetic screens, clinical exam results, patient history, and clinical diagnosis for AD. Following each visit, patients are assigned one of the three diagnostic categories: Cognitively Normal(CN), Mild Cognitive Impairment(MCI), and Alzheimer's Disease(AD).

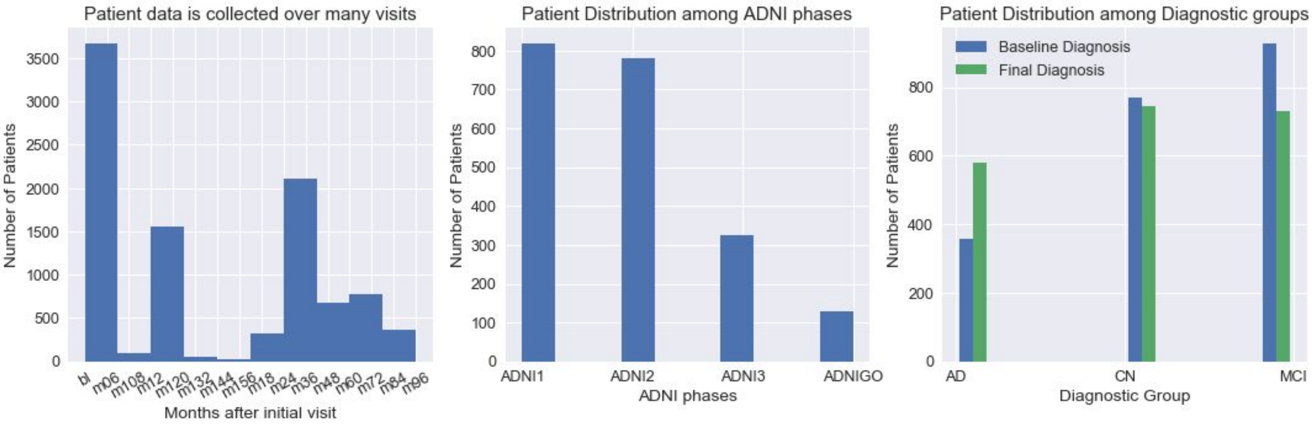

Figure 1: ADNI contains longitudinal data collected over multiple phases across many patient visits.

ADNI provides a dataset (adnimerge) that summarizes the most commonly used measures captured throughout the study. This dataset compiles 113 features about more than 13,000 visits and 2,080 patients and formed the initial basis of our project data. Of the 2080 patients in adnimerge, only 2056 ever received a diagnosis after the baseline visit. The data comes from roughly an equal distribution of male and female patients aged between 54 and 91 years old with a mean age of 73 years old. The education level of the patients follows a roughly equal distribution across the three diagnostic categories in the data. It is noteworthy that the distribution of race and ethnicities in the data, however, is largely skewed towards white and non-hispanic populations (Figure 2).

Next, we expanded our feature set to include features outside of adnimerge. Many of the features in the merged data such as Amyloid-β and Tau proteins are well-studied correlates of Alzheimer's disease. Although these features are likely to improve the predictive power of our Alzheimer's diagnosis model, we also wanted to include features that are not commonly studied to allow for the possibility of revealing surprising or lesser correlates of Alzheimer's disease. We cleaned and aggregated an additional per-patient dataset from a subset of the raw data by manually selecting features datasets containing information on patient demographics, medical history, neurological exams, questionnaires, and biomarkers. Ultimately we chose 13 datasets for completeness and consistency across ADNI phases. Even within this subset, the raw datasets we chose showed considerable variation in the number of records per patient, formatting of visit codes, and missingness. We first removed features that were irrelevant (e.g. metadata), dependent on other features, or too complex to analyze (e.g. doctor's notes), and imposed standard data types and missing value indicators. Visit code formatting in the adnimerge dataset can be used to construct a per patient baseline dataset by extracting the baseline observation for each patient. Unfortunately, the same strategy is impossible in the raw data due to inconsistent visit code formatting across raw datasets. Therefore, we recorded the data from the first observation belonging to each patient to approximate a per patient baseline measurement from the raw data and combined the resulting data with adnimerge. The combined, cleaned dataset includes measures from 257 features recorded from more than 2500 patients in total.

Data preprocessing methodology

Before proceeding to the description of our preprocessing strategy, it is important to highlight that, given the course's timeframe, the group has decided to leave imaging data out of our scope.The group was initially interested in understanding the key factors leading to the development of Alzheimer's Disease. Because of the longitudinal nature of collection, we encountered a number of preprocessing challenges before we could address our primary research question.

A) Per patient data aggregation: Due to the longitudinal nature of data collection, the datasets that we encountered were organized by the visitcode under which a subset of the features were collected for each patient. Furthermore, the raw data was collected across multiple ADNI phases (Figure 1) with some features and patients present in only one or two phases. For both the convenience of modeling and also to reduce the prevalence of missingness in the data, we converted the data to a per patient format by extracting the baseline observation for each patient. Due to inconsistent diagnostic classes among ADNI phases, we also converted patient final diagnosis (our response variable) to the three most commonly available categories: CN, MCI, AD. Therefore, we first described the raw data in a per-patient form such that each data point corresponded to a single patient entry. In total, this process collapsed the dataset of 13,000 visits to one with 2056 patients.

B) Missing data: Another central challenge in working with the datasets provided by ADNI is the number of missing data caused by lack of repeated measurements across the different phases, patient discontinuities and temporal design of the study. (Figure 3). A view of the raw per-patient dataset shows examples of the data missing at random and large non-random chunks. Similarly, a look per feature in the data shows that more than 50 out of the 113 features on adnimerge have more than 50% of missing entries (Figure 4).

We dealt with the problem of missing data in the following ways:

1. Features in the dataset that were completely void of information were dropped.
2. Patient observations without diagnosis codes were removed.
3. For other missing features in the baseline dataset, we wrote a function that selected the first non-null value from subsequent visits, given that the diagnosis code had not changed from the baseline diagnosis. This approach was able to fill about 4000 of the 40743 missing values in the per-patient dataset.
4. For filling the rest of the dataset, we applied model based imputation approaches. Before imputing missing values, we standardized the data with the reason that imputing before standardization may cover up or dilute any bias present in the data.

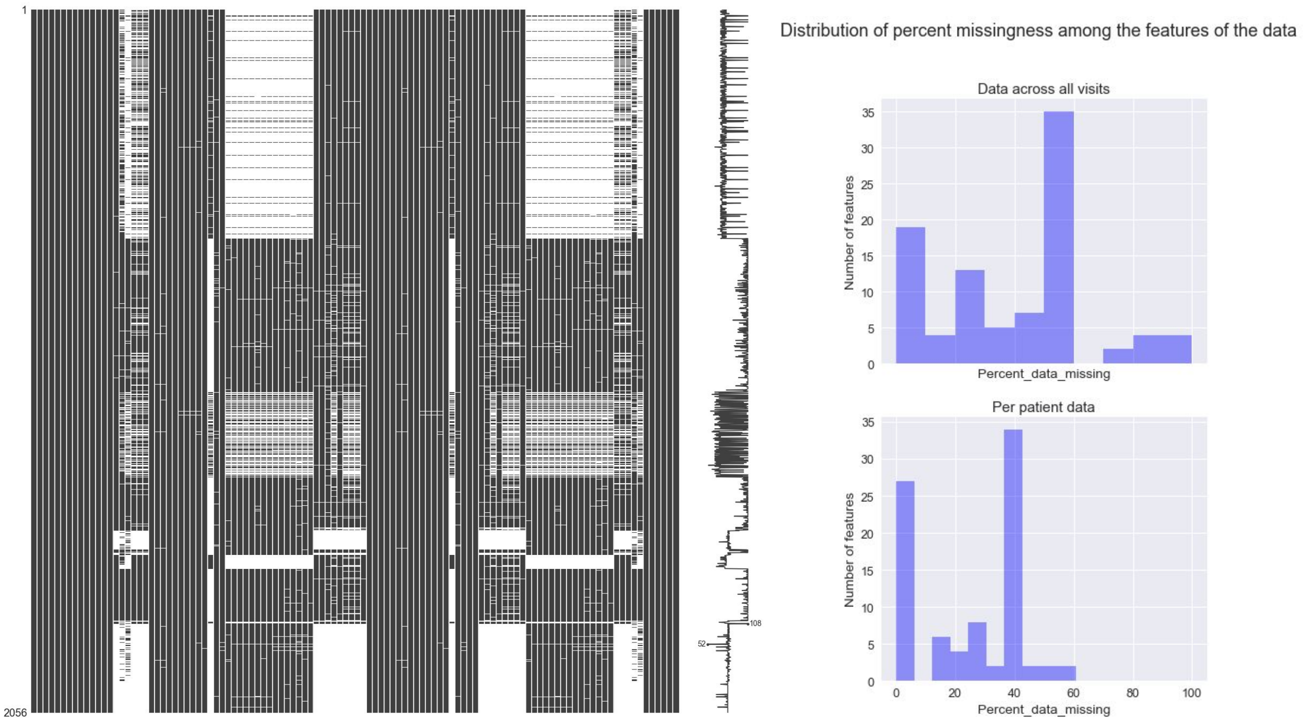

Figure 3: Left Panel: Missing data overview of the adnimerge dataset. White indicates missing values. Sparkline on the right represents data completeness by row and the least(52) and highest(108) number of features missing in each entry of the database. Right Panel: The distribution of percent missing among the features present in the dataset.

C) Highly Correlated Features: As seen by the distribution of pairwise correlation between features, another major issue with the dataset is the collinearity among the features. This primarily occurs because there are repeat measurements of the same feature across multiple visits for the same patient. To ameliorate this problem, we computed such pairwise correlation between all features in the dataset and removed all the features which were correlated with one another with a correlation coefficient of greater than 0.95. This approach reduced the number of features in the dataset to 64 from 113. We also removed those features that were missing across >98% of patients.

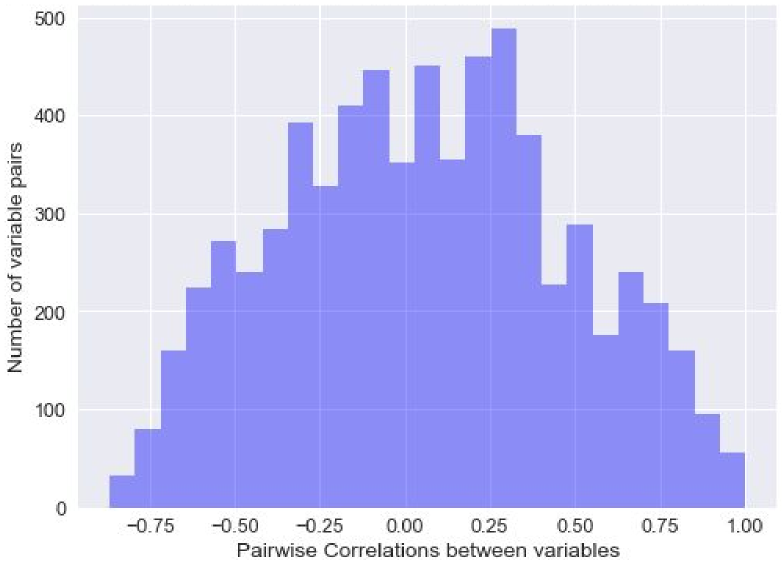

Figure 4: Distribution of correlation between pairs of features in the merged data

At this point in the process, we have a cleaned dataset in which we have:

1. Removed highly correlated, duplicated or informationless features.
2. Removed patients with no diagnosis decision
3. Filled missing values with existing information within the collected data or model based imputation

Preliminary modeling results

We used the resulting dataset to run initial modeling to check the consistency of our imputation. We fit a regularized logistic regression with an L2 penalty using cross validation to find an optimal regularization parameter. Our aim was to identify the variables that emerged as the best predictors of the cognitive state (CN, MCI, and AD) of ADNI patients. The model showed some predictive power but was overfit to the training data, with training and test accuracies of 83% and 73% respectively. Visualization of the strongest predictors of patient diagnosis (Figure 5) suggest that we need further investigation into our imputation strategies during the upcoming stages of the Final Project. While some predictive features like clinical dementia rating (CDRSB) match our expectations of being a good predictor for AD and MCI, others do not. Some unexpected findings include a negative correlation between the concentration of amyloid-β protein and age to AD.

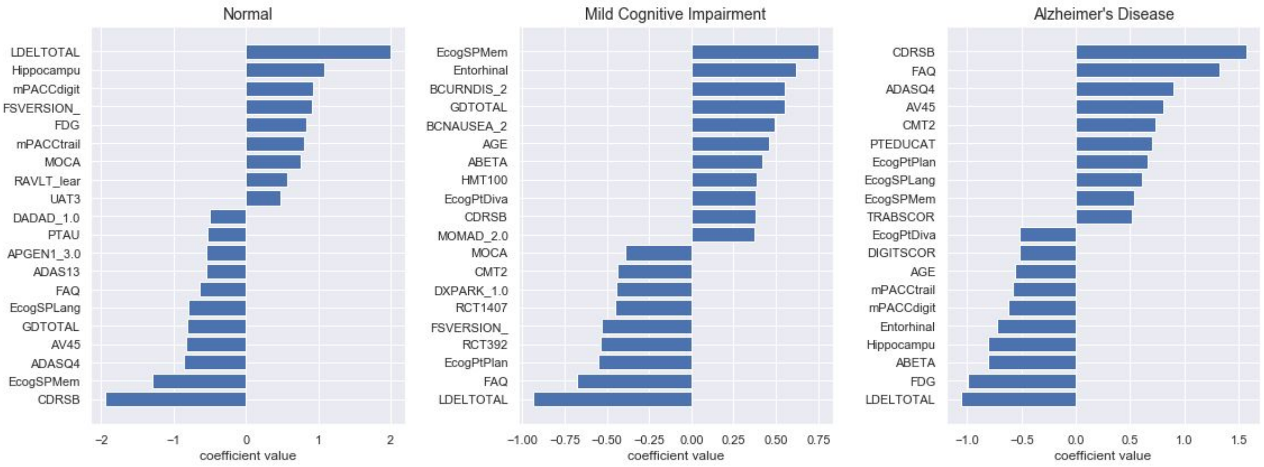

Figure 5: Coefficients from initial regularized logistic model

These unexpected results may be linked with the high correlation between features, problems in our imputation strategy, flaws in our modelling strategy or other unseen issues. Those are all matters that we want to further investigate on the last part of our project.

Revised Research question

After the initial exploratory analysis of the data, we have decided to address the following question on our Final Project: *“What are the key clinical, genetic, and biospecimen-related features that can predict whether an individual has developed or will develop Alzheimer’s Disease at later point in time? How can different data imputation strategies impact predictions?”*