Final Project Proposal

1. The Team

A list of the members of your team and their expected roles.

Edgar Lorenzo:
Kristine Carnavos:
Rahul Sinha: Data Analyst
Mary Mulrooney:

2. **Problem Statement and Background (**Business Understanding)

What are the business needs you are trying to address? What are the objectives of the data project? For example, if you are at a telecommunications company, that needs to retain its customers, can you build a model that predicts churn? Maybe you are interested in using live data to help better predict what coupons to offer what customers at the grocery store.

Alzheimer's disease is the most common form of dementia, affecting over 5.7 million Americans and 10% of all adults ages 65 and older. Currently, there is no cure for Alzheimer's Disease, only therapies that help to lessen the severity of symptoms and improve the patient's quality of life. A formal diagnosis is only achieved currently via extensive cognitive testing, blood testing, and expensive brain-imaging using MRI scans. There is no one diagnostic tool that can definitively diagnose Alzheimer's itself, and currently, there is no way to diagnose Alzheimer's disease prior to symptoms occurring.

MRI scans are currently being used to gather data in hopes of developing an understanding of possible markers of the disease even before symptoms begin to occur. This will allow for early intervention on patients identified to have the most critical input variables as well as targeting them for therapies and medications that could help to lessen the severity of the disease's impact on the patient's lives. It will also help to lessen the costs of Alzheimer's diagnostics for patients as well as free up resources for medical institutions that would otherwise be tied up in helping to diagnose Alzheimer's cases.

We work for a pharmaceutical company that has just discovered a drug that works to prevent Alzheimer's disease from progressing into later stages. We aim to produce a model that will be able to predict if a patient will develop Alzheimer's disease so we can provide this drug before their symptoms worsen. This will also help us to justify the expense to insurance companies as well as patients and physicians.

3. The Data Source(s) You Intend to **Use**

Describe the data source(s) you will use. Make sure you have access to the data you want to use *in the quantity and quality you need*. Describe how you plan to obtain the data, or how you got it if you already have it. Give a summary of the cleaning/joining of data that you expect to do up front.

The following datasets were obtained from Kaggle.com and come from the Open Access Series of Imaging Studies (OASIS) project. The datasets are overall complete but some cleaning is necessary as some attributes have empty cells. These attributes include EDUC, SES, MMSE, and CDR. It seems that these individuals did not have dementia so the cells were left blank. However, for SES it is possible that the data was unable to be obtained. In the case of CDR, it is possible to insert 0's for individuals who do not have dementia. This practice can carry over into other attributes except for SES. With this attribute, it is best to delete the lines for which SES is not available.

Oasis-Cross-Sectional.csv

Oasis-Cross-Sectional.csv

Rows: 436

Columns: 12

Attributes:

ID – Categorical

Subject unique Identifier

M/F – Categorical

Gender(M/F)

Hand – Categorical

Dominant Hand(All R)

Age – Numerical

Age in years(18-96)

Educ – Numerical

Education Level(1-5)

SES – Numerical

Socioeconomic Status(1-5)

MMSE – Numerical

Mini Mental State Examination(14-30)

CDR – Numerical

Clinical Dementia Rating(0-2)

eTIV – Numerical

Estimated Total Intracranial Volume(1.12K-1.99K)

nWBV – Numerical

Normalize Whole Brain Volume(.64-.89)

ASF – Numerical

Atlas Scaling Factor(.88-1.56)

Delay – Categorical

Rows: 373

Columns: 15

Subject ID – Categorical

Subject identification

MRI IDMRI – Categorical

Exam Identification

Group – Categorical

Class

Visit – Numerical

Visit Order

MR Delay – Numerical

MR Delay Time (Contrast)

M/F – Categorical

Hand – Categorical

Age – Numerical

EDUC – Numerical

Years of Education

SES – Numerical

Socioeconomic Status

MMSE – Numerical

Mini Mental State Examination

CDR – Numerical

Clinical Dementia Rating

eTIV – Numerical

Estimated total intracranial volume

nWBV – Numerical

Normalize Whole Brain Volume

ASF – Numerical

Atlas Scaling Factor

4. Goals of Your Analysis

List some goals of your analysis, ideally in the form of a testable hypothesis, or via well-defined success metrics. These can be tentative, and you don't need to stick to them throughout your project. Again since you haven't done any exploratory analysis yet, you might assume that the data has a structure that it doesn't, and you might not have seen other interesting patterns in the data. However, it would help if you always approached the data with some expectations so that your efforts are focused.

Our goal is to determine which variables, if any, have strong impact to predict our target output variable- CDR rating. We can then create a linear regression model with those outputs to see how much of the variation can be explained.

Specifically, it would be interesting to see if education level and socioeconomic status have a link to developing Alzheimer's disease and what the variation is. It would also be of interest to health care providers to see how the Mini Mental State Examination (MMSE) is correlated to dementia. This is a 30-point questionnaire that is a quick, simple, and costless exam that all providers can use to test their patients. If this is an indicator of dementia, perhaps we can lobby that all physicians must perform this exam annually for their patients. This could allow for early detection and early treatment with our drug (and more patients we are able to treat for a longer period of time).

5. Description of Data Analysis Tools You Plan to Use

Describe the tools you plan to use throughout the project. As you might expect, there will be several stages in the project (SEMMA). This part can also be tentative, and we will give you feedback on your analysis plan as part of grading the assignment.

We will follow the different SEMMA steps in our project.

1. Purpose: predict our target output variable- CDR rating.
2. Sample: We will use the data from Oasis-Cross-Sectional.csv. The dataset is small with 436 rows and hence we do not need to sample from the data.
3. Explore: In this step, we will explore, clean, and preprocess the data. Examine the dataset using JMP's Analyze-Distribution, Boxplots, Graph Builder. Identify the types of variables as the three JMP modeling types- Continuous, Nominal, Ordinal. Use the Analyze > Screening tools to explore outliers and missing values. We will use the Multivariate Correlations table to explore the correlation among predictors. The dataset has 436 rows and 12 columns, hence, Dimensionality should not be an issue. But we will do a Principal Component Analysis to examine the columns.
4. Modify: Based on the explore step, we may want to transform the variables and impute missing values using multivariate normal imputation or ignore records with missing values. Based on the JMP Pro Explore Outliers analysis and our domain knowledge, we would want to correct the values of some outliers or exclude the records with the outliers from our model. Additionally, using JMP's Column Recode tool we may have to recode the Delay column as well as other nominal columns. Finally, we use JMP's "Make Validation Column" utility to partition data into training and validation datasets, which we will use to build and validate our prediction model.
5. Model: Fit predictive models. After creating the training and validation datasets, we use the "Fit Model" option in JMP to fit Regression models.
6. Assess: Compare models using a validation dataset and JMP tools such as Crossvalidation, Confusion Matrix, and Distribution to summarize errors for training and validation sets. Assess the MAD and MAE of different models.

6. Describe the Data Products Your Project Will Produce

Data products include results of statistical tests, performance analyses of learning algorithms, visualizations of the data or model parameters. Give a list of business insights to your project.

Data products:

1. Hypothesis tests to support claims

2. Association mining results
3. Performance analysis using crossvalidation, confusion matrix, graphs, tabulate, etc.
4. Data visualizations through Graph Builder for easy representation of trends and patterns
5. The regression model with formula, error report, distribution of residuals
6. Our model will identify predictors for dementia

We aim to produce a model that will be able to predict if a patient will develop Alzheimer's disease so we can provide this drug before their symptoms worsen. The model will help develop an understanding of possible predictors of the disease even before symptoms begin to occur. This understanding can help in the early detection of patients thus reducing the costs of Alzheimer’s diagnostics for patients as well as freeing up resources for medical institutions that would otherwise be tied up in helping to diagnose Alzheimer’s cases.
The above use cases will also help us to justify the expense to insurance companies as well as patients and physicians.

7. Reference

Please list all references that will be used for this project.

1. DATA MINING FOR BUSINESS ANALYTICS: CONCEPTS, TECHNIQUES, AND APPLICATIONS WITH JMP PRO® - Shmueli, Bruce, Stephens, Patel
2. Kaggle.com