Alzheimer's Disease Prediction

Tessa Clary
Applied Computer Science
University of Colorado
Boulder, CO
sccl8634@colorado.edu

Problem Statement and Motivation:

Alzheimer's Disease is a complex and devastating condition that impacts millions of people around the globe. Early detection is paramount for patient care as well as advancing our knowledge of the disease. The motivation for this project is to develop a predictive model that can help identify those who are at risk of developing early-stage Alzheimer's. This will be done using the techniques and knowledge that we have learned about data mining and analysis in this course. The goal of this project is to discover key risk factors and patterns that can help lead to earlier diagnosis and better treatments.

Literature Survey:

"Multivariate Logistic Regression Analysis of Complex Survey Data with Application to BRFSS Data"
This study uses a multivariate logistic regression analysis to study the association of risk factors for heart attacks and strokes. This closely aligns with the goals of this project albeit with a different focus.

"A smart Alzheimer's patient monitoring system with IoT-assisted technology through enhanced deep learning approach"
This study utilized IoT devices and deep learning techniques to help predict Alzheimer's. Given an accuracy and precision rate of 98% and 97% respectively, there is obviously a lot of value in using IoT to accurately monitor patients in real time. This sets a high benchmark for this project as similar rates of accuracy and precision should still be potentially achievable.

"Population measures of subjective cognitive decline: A means of advancing public health policy to address cognitive health"
This study focuses on SCD, or subjective cognitive decline. It is an early cognitive indicator that complements this project given that it includes an early-stage assessment parameter. Given the issues with studying Alzheimer's Disease via population-based measures, this can potentially help alleviate the difficulty.

Proposed Work:

The data set will be collected and cleaned, with inconsistencies, missing or incorrect values, and outliers being identified and handled. Relevant data from the literature survey will also be examined for pertinence; integrating SCD is likely.

After this, data preprocessing will begin. Feature engineering will be used to extract relevant variables, and the data will be normalized for consistency. Data visualization here can also help identify anomalies or patterns. Data transformation will then be carried out at this point in the form of finalizing the solution to issues such as missing values or converting data types.

Given that this project is focused on finding early-stage Alzheimer's Disease, the model design will focus on demographic data, health factors, and relevant correlations in an attempt to predict patient outcomes. To assess the model's

performance, a number of evaluation methods will be used including accuracy, precision, and recall.

The difference between this project and prior works is that the given prior works had a far narrower scope of predictive factors. This will be a more inclusive and comprehensive analysis of overall demographic and health factors, which may dilute the model's predication capability, but should also do a better job of showing which variables if any impact the development of the disease.

Data Set:

This project focuses on a published data set from the CDC called: "Alzheimer's Disease and Healthy Aging Data". It is 250,000 rows with 39 column labels. It includes variables for:

- Year Start / End
- Location
- Disease Indicator Category:
    - Caregiving
    - Cognitive Decline
    - Mental Health
    - Nutrition/Physical Activity/Obesity
    - Overall Health
    - Screenings and Vaccines
    - Smoking and Alcohol Use
- Age Group
- Gender
- Race

Data.gov Dataset:
https://catalog.data.gov/dataset/alzheimers-disease-and-healthy-aging-data

Same Dataset hosted by the CDC:
https://data.cdc.gov/Healthy-Aging/Alzheimer-s-Disease-and-Healthy-Aging-Data/hfr9-rurv

CDC's page that uses the same dataset for visualization purposes:
https://www.cdc.gov/aging/agingdata/index.html

Evaluation Methods:

This dataset will be evaluated primarily through the use of a regression model- accuracy, precision, recall, F1 score, and ROC-AUC will be considered as metrics. A confusion matrix will also reveal useful insights into the model's performance as well. In the event that the initial model is too generalized to return quality performance metrics, additional data transformation may be carried out alongside other scope changes in order to ensure a final model with a real capacity to predict. Additional models may be developed beyond the initial regression as needed to further explore the data set.

Tools:

The majority of the code for this project will be written in Python for its flexibility with data manipulation, analysis, and modeling. Numpy is also going to be potentially used given its ability to handle numerical operations and arrays. Pandas is irreplaceable when it comes to data manipulation, cleaning and structuring so it will be relied on heavily. SciPy expands the scientific and technical computing functions available to the project, so it may come into play if the initial analysis does not have quality results. Scikit-Learn provides additional ML algorithms and tools for pertinent use as well. SQL may be used depending on how easy the data set is to load and manipulate. For data visualization, Matplotlib will be used with the potential for either Seaborn or Plotly as needed. Github will be used for version control and project management.

Milestones:

1. Data Preprocessing (1 Week)
    a. The data set will be cleaned, with any issues or gaps in it addressed.
    b. Exploratory data analysis will be performed to gain early insights.
    c. Preprocessing will take place during this time, including feature engineering, normalization, and data transformation.
2. Model Design (1 Week)
    a. During this stage, the scope of the initial model will be defined and developed, with its objectives and target variables specified.
    b. Model architecture will be chosen and a baseline model will be developed.
3. Model Training (1 Week)
    a. The model will be trained using the preprocessed data.
4. Evaluation (1 Week)
    a. Its performance will then be benchmarked against other relevant solutions to help gauge its relative accuracy alongside other evaluation metrics.
    b. Accuracy, precision, recall, F1 score, and ROC-AUC along with other measures as needed.
5. Model Optimization (1 Week, Project Part 3 Goal on 12/4)
    a. Based on the results of the evaluation, additional tuning and refinement of the model may be necessary to a variable degree at this stage.
6. Write-up (1 Week)
    a. Documentation will be finalized, and project conclusion will have been reached.

Milestone 1 is currently being carried out, with milestones 2 and 3 to be performed over the next few weeks. By fall break, milestones 3 and 4 will be in progress. At this stage, initial analysis will likely be complete and the feasibility of the project goal can then be reexamined if the scope likely needs to be updated. Ideally this allows for focus on model optimization for the rest of the course so long as the model does not require extensive rework. Once the optimizations are sufficient, the remainder of the available time will be spent preparing the project presentation.

Milestones Completed:

1. Data Preprocessing
    a. The data was cleaned, with irrelevant or duplicate columns removed.
    b. Rows with missing data points were discarded.
    c. Rows that did not include the pertinent variables were removed as well.
    d. Dummy variables were implemented for the qualitative age and race independent variables.
    e. A dummy variable was also created to represent the reporting of subjective cognitive decline by an individual.
2. Model Design
    a. The original model goal had been for linear regressions initially. Once the dataset was more thoroughly examined and worked with, it was found that dummy variables needed to be implemented.
    b. Due to the Boolean nature of dummy variables, the model

design was updated to be a logistic regression in order to better accommodate the qualitative variables.

3. Model Training
    - **a.** Both linear and logistic regression models were separately created for both the age and race independent variables using the preprocessed data.
    - **b.** The dataset was temporarily split here for ease of use between the two models- it will be revisited and updated once the model has been finalized.
4. Evaluation
    - **a.** Evaluation metrics were added to the models at this stage.
    - **b.** The linear regression models are being dropped at this point due to irrelevancy.

Milestones Todo:

1. Model Optimization
    a. The set of independent variables needs to be finalized and the dependent variable of subjective cognitive decline (SCD) needs to be reexamined as the nature of the dataset does not lend itself readily to a regression without a fair amount of data transformation; refinement of this aspect should heavily improve the model.
    b. Other models beyond regressions should be considered at this stage due to a possibility of a better fit for the data with less needed transformation.
2. Write-up
    a. The abstract will need to be updated based on any model updates and their new results.
    b. A proper separate introduction will be added to the report.
    c. A discussion on the data set and its details as well as what I did with it will also be included.
    d. Main techniques used in the project and key results from the final model analysis will be discussed.
    e. The paper will finish with a discussion on the applications of the knowledge that was found over the course of the project.
    f. Visualizations of the results will be created provided enough time after model completion.

Results so far:

This dataset has presented far more challenge to work with than I had initially expected when selecting it. The Behavioral Risk Factor Surveillance System (BRFSS) is a telephone survey run by the CDC since 1984. This dataset in particular contains the relevant data points from BRFSS for Alzheimer's Disease as well as on the types of healthy aging the participants engage in from 2015 to 2021. The main difficulty in working with this dataset so far is the fact that the primary numerical metric used for gauging response to a specific question is represented as a percentage that cannot be easily related to a variable. Between this and the fact that those suffering from SCD are categorized under one topic makes it difficult for me to understand how to relate other potentially related topics without such an important shared metric. As a result, the scope of the project changed to focus on the independent variables that can be successfully related to those reporting SCD- either age, race, or gender was reported alongside each entry. This led to

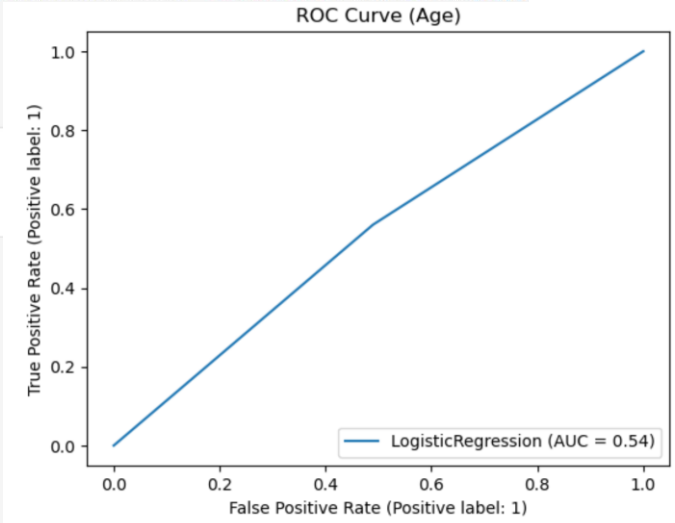
the splitting of the dataset, so that each set of independent variables could be examined without their respective data being influenced by rows that do not hold relevancy.

At this stage a complete rework of the model is being considered; there may be models other than regressions that could be a better fit for the dataset, but I am not sure what could work better that would come to a stronger conclusion. The literature survey section was updated to include a very relevant Data Science paper using BRFSS to perform a logistic regression- unfortunately the paper is well beyond my scope of implementation, as I am unsure of how to properly manipulate my dataset to also show such viable results. Going forwards age, gender, and race will be finalized as separate models to read into their relevancy to SCD before going forward with my best attempt at a multivariate model that tries to encompass as many relevant independent variables as possible.

Regressions on Age:

```python
#Logistic Regression on Age
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix
import matplotlib.pyplot as plt
from sklearn.metrics import plot_roc_curve
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)

file_path = 'alzage.csv'
data = pd.read_csv(file_path, header=0)

dependent_variable = 'cog_decline'
independent_variables = ['50-64 years', '65 years or older']

x = data[independent_variables]
y = data[dependent_variable]

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=50)

model = LogisticRegression()
model.fit(x_train, y_train)

predictions = model.predict(x_test)

accuracy = accuracy_score(y_test, predictions)
print("Accuracy:", accuracy)

feature_importance = pd.DataFrame({'Feature': independent_variables, 'Coefficient': model.coef_[0]})
print(feature_importance)

plot_roc_curve(model, x_test, y_test)
plt.title('ROC Curve (Age)')
plt.show()
```

```
Accuracy: 0.7558488302339532
              Feature  Coefficient
0        50-64 years    -0.004645
1  65 years or older     0.004789
```



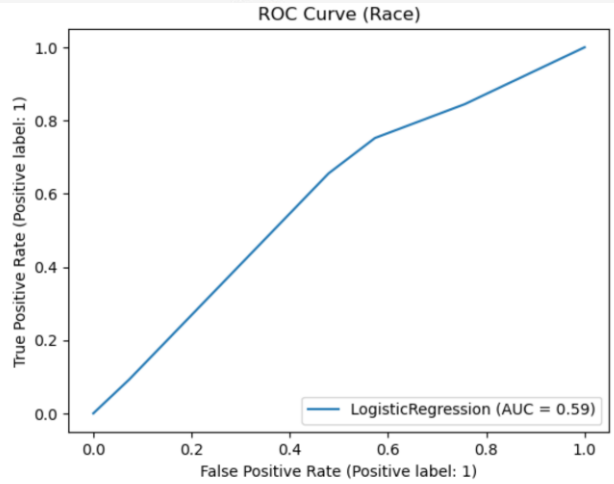

```python
#Linear Regression on Age
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error

file_path = 'alzage.csv'
data = pd.read_csv(file_path, header=0)

dependent_variable = 'cog_decline'
independent_variables = ['50-64 years', '65 years or older']

x = data[independent_variables]
y = data[dependent_variable]

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=50)

model = LinearRegression()
model.fit(x_train, y_train)

predictions = model.predict(x_test)

r_squared = r2_score(y_test, predictions)
print("R-squared:", r_squared)

mse = mean_squared_error(y_test, predictions)
print("Mean Squared Error:", mse)

feature_importance = pd.DataFrame({'Feature': independent_variables, 'Coefficient': model.coef_[0]})
print(feature_importance)
```

```
R-squared: 0.00023805242345531497
Mean Squared Error: 0.18449744554611705
              Feature  Coefficient
0        50-64 years    -0.000871
1  65 years or older    -0.000871
```

Regressions on Race:

```python
#Logistic Regression on Race
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix
import matplotlib.pyplot as plt
from sklearn.metrics import plot_roc_curve
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)

file_path = 'alzheimer_data_copy_race.csv'
data = pd.read_csv(file_path, header=0)

dependent_variable = 'cog_decline'
independent_variables = ['Asian/Pacific Islander', 'Native Am/Alaskan Native', 'Black, non-Hispanic', 'Hispanic', 'White, non-Hispanic']

x = data[independent_variables]
y = data[dependent_variable]

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=50)

model = LogisticRegression()
model.fit(x_train, y_train)

predictions = model.predict(x_test)

accuracy = accuracy_score(y_test, predictions)
print("Accuracy:", accuracy)

feature_importance = pd.DataFrame({'Feature': independent_variables, 'Coefficient': model.coef_[0]})
print(feature_importance)

plot_roc_curve(model, x_test, y_test)
plt.title('ROC Curve (Race)')
plt.show()
```

```
Accuracy: 0.9833578071463533
                    Feature  Coefficient
0    Asian/Pacific Islander     0.381099
1  Native Am/Alaskan Native    -0.030640
2       Black, non-Hispanic    -0.394249
3                  Hispanic    -0.296357
4       White, non-Hispanic     0.340152
```




```python
#Linear Regression on Race
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error

file_path = 'alzheimer_data_copy_race.csv'
data = pd.read_csv(file_path, header=0)

dependent_variable = 'cog_decline'
independent_variables = ['Asian/Pacific Islander', 'Native Am/Alaskan Native', 'Black, non-Hispanic', 'Hispanic', 'White, non-Hispanic']

x = data[independent_variables]
y = data[dependent_variable]

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=50)

model = LinearRegression()
model.fit(x_train, y_train)

predictions = model.predict(x_test)

r_squared = r2_score(y_test, predictions)
print("R-squared:", r_squared)

mse = mean_squared_error(y_test, predictions)
print("Mean Squared Error:", mse)

feature_importance = pd.DataFrame({'Feature': independent_variables, 'Coefficient': model.coef_[0]})
print(feature_importance)
```

```
R-squared: 0.002156712477616618
Mean Squared Error: 0.016329935174344806
                    Feature   Coefficient
0    Asian/Pacific Islander  5.494076e+09
1  Native Am/Alaskan Native  5.494076e+09
2       Black, non-Hispanic  5.494076e+09
3                  Hispanic  5.494076e+09
4       White, non-Hispanic  5.494076e+09
```