SAS® GLOBAL FORUM 2020

MARCH 29 – APRIL 1
WASHINGTON, DC

USERS PROGRAM

A Data-driven Approach to Understand Mild Cognitive Impairment from Imbalanced Data Using SAS®

Liyuan Liu[1], Meng Han[2*], Yiyun Zhou[1], Gita Taasoobshirazi[3]

1. Analytics and Data Science, Kennesaw State University
2. College of Computing and Software Engineering , Kennesaw State University
3. Department of Statistics and Analytical Sciences , Kennesaw State University

Abstract

Cognitive decline has emerged as a significant threat to both public health and personal welfare, and mild cognitive decline/impairment (MCI) can further develop into Dementia/Alzheimer's disease. While treatment of Alzheimer's disease can be expensive and ineffective sometimes, the prevention of MCI by identifying modifiable risk factors is a complementary and effective strategy. Using a data-driven approach to understand the MCI factors become a crucial research question recently. However, there is a main problem that **most healthcare datasets are imbalanced**. Therefore, we employed **multiple strategies to deal with imbalanced data**, such as random oversampling, random under-sampling, SMOTE, SMOTEENN, etc. After that, to examine the effects of comparing multiple strategies and different machine learning algorithms, we use three machine learning algorithms: **decision tree (DT), neural networks(NN), and Gradient Boosting (GB)**. In this study, we not only to compare different balanced strategies and machine learning algorithms but also investigate the most important factors that contribute to MCI.

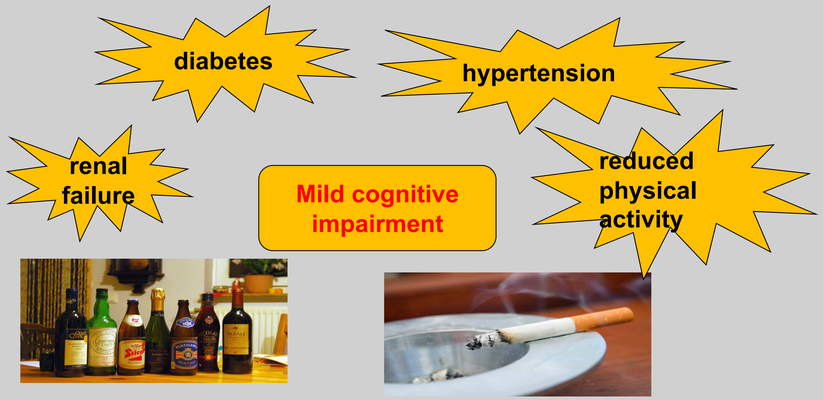


Fig 1: The MCI Factors Reported by Previous Research

Introduction

Alzheimer's is a type of dementia that **causes problems with memory, thinking, and behavior.** Symptoms usually develop slowly and get worse over time, becoming severe enough to interfere with daily tasks. Mild cognitive impairment (MCI) causes **a slight but noticeable and measurable decline in cognitive abilities, including memory and thinking skills.** A person with MCI is at an increased risk of developing Alzheimer's or another dementia.

However, **between 2002-2012, 99% clinical trials for the treatment of Alzheimer's disease failed.** There are several limitations of previous research: 1. Rely on well-controlled lab experiment and clinical conservation, which is time and resource-consuming 2. A limited number of factors studied.

Therefore, we proposed **a data-driven approach to re-exam MCI factors.** To implement machine learning algorithms to predict MCI, the most challenge we meet is the **highly imbalanced data**. We employed five different balanced strategies to address the imbalanced problem. In the results of our experiments, our best strategy **increased recall from 0.007 to 0.85**. In this study, we found that depression, physical health, cigarette usage, education level, and sleep time play an important role in cognitive decline, which is consistent with the previous discovery. Besides that, the first time, we point out that other factors such as arthritis, pulmonary disease, stroke, asthma, marital status also contribute to MCI risk, which is less exploited previously.

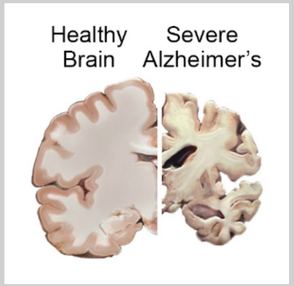


Fig 2: Brian Comparison Between Healthy Brain And Severe Alzheimer's

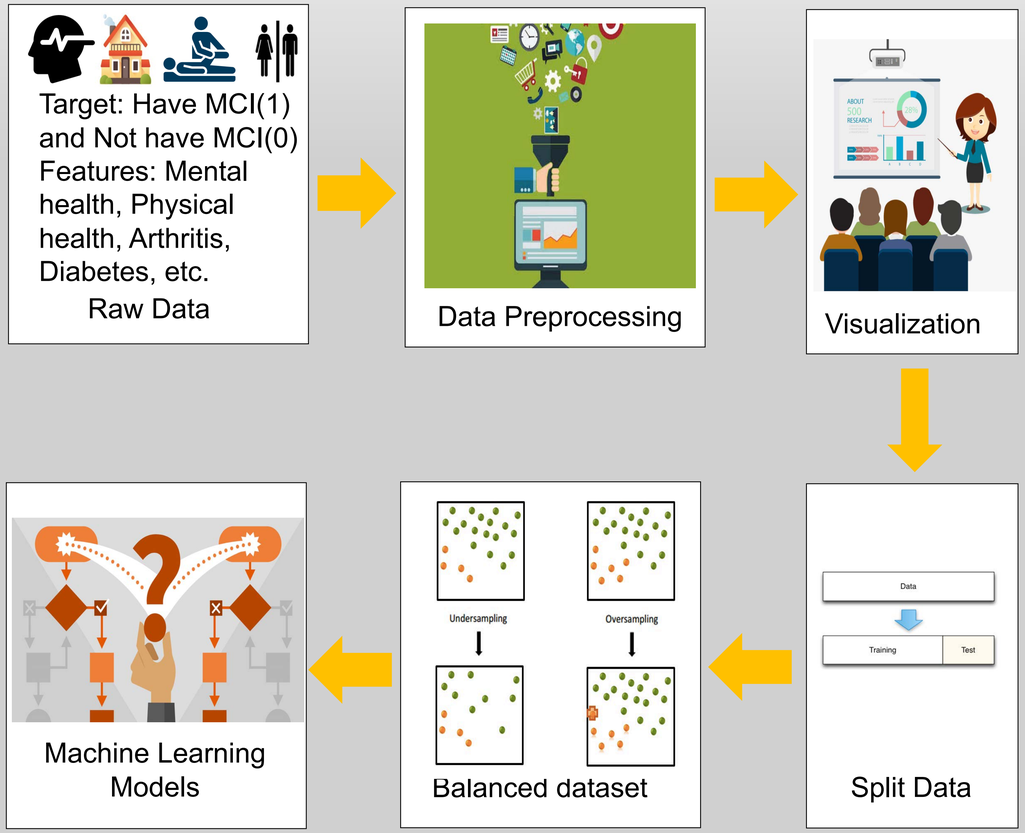
A Data-driven Approach to Understand Mild Cognitive Impairment from Imbalanced Data Using SAS®

Liyuan Liu[1], Meng Han[2*], Yiyun Zhou[1], Gita Taasoobshirazi[3]

1. Analytics and Data Science, Kennesaw State University
2. College of Computing and Software Engineering , Kennesaw State University
3. Department of Statistics and Analytical Sciences , Kennesaw State University

Method

Fig 3: Workflow of this study



The dataset collected from Centers for Disease Control and Prevention (CDC), there includes 32 features and 1 target (binary) variable. There are total 60816 observations in the dataset.

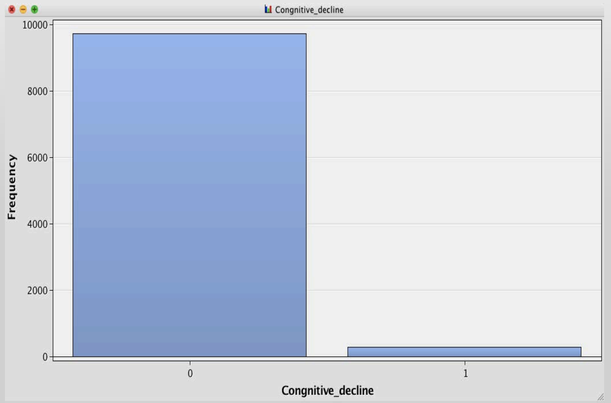


Fig 4:Distribution of Target Variable

Extremely Imbalanced Data!!!!

To solve the imbalanced-class problem, we employ several strategies:

- Random Over-sampling
- Random Under-sampling
- SMOTE(Advanced over-sampling)
- SMOTEENN(Advanced combine over-sampling and under-sampling)
- SMOTETomek (Advanced combine over-sampling and under-sampling)

To predict MCI and find the feature importance, we employ several machine learning algorithms:

- Decision Tree.
- Gradient Boosting.
- Neural Network.

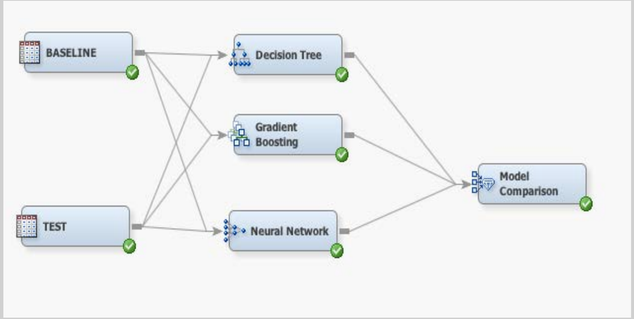


Fig 5: Example Workflow of model Comparison

A Data-driven Approach to Understand Mild Cognitive Impairment from Imbalanced Data Using SAS®

Liyuan Liu[1], Meng Han[2*], Yiyun Zhou[1], Gita Taasoobshirazi[3]

1. Analytics and Data Science, Kennesaw State University
2. College of Computing and Software Engineering , Kennesaw State University
3. Department of Statistics and Analytical Sciences , Kennesaw State University

Results

In MCI detection: **Recall** is more important:
It is obviously important to catch every possible MCI even if it means that the authorities might need to go through some false positives.

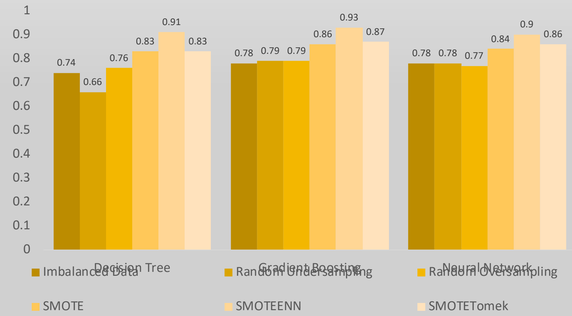


Fig 6: ROC AUC Comparison by Different Balanced Strategies

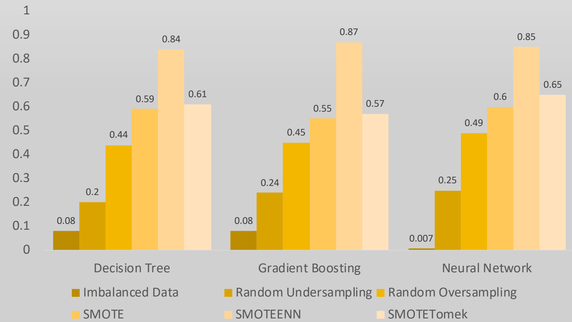


Fig 7: Recall Comparison by Different Balanced Strategies

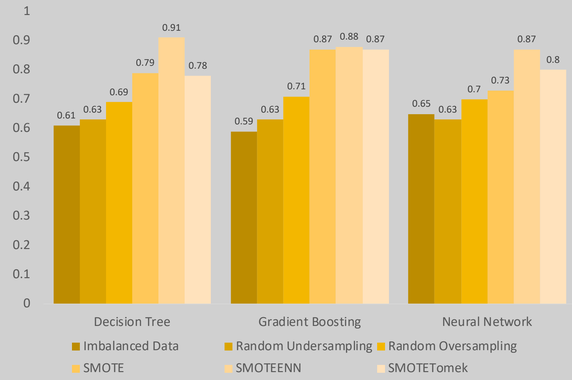


Fig 8: Precision Comparison by Different Balanced Strategies

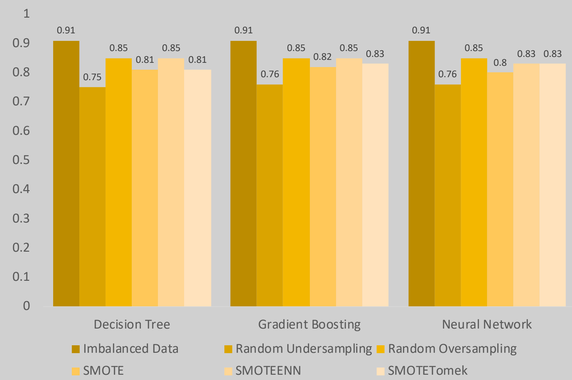


Fig 9: Accuracy Comparison by Different Balanced Strategies

Results

Balanced Strategies increased recall from 0.007 to 0.87!

Variable Name	Importance ▼
Mental_health	1
Married	0.728181
Education_Level	0.676788
Physical_health	0.643391
Exercise	0.633698
Divorced	0.45699
Widowed	0.421736
Never_Married	0.3535
could_not_see_doctor	0.289047
Rent	0.269693
Own	0.254646
number_of_children	0.208545
Sleep_time	0.176987
Smoking	0.170275
Dental	0.146773
A_member_of_unmarried_couple	0.135645
skin_cancer	0.111733
Asthma	0.062876
Depressive_disorder	0.020423

Fig 10: Feature Importance

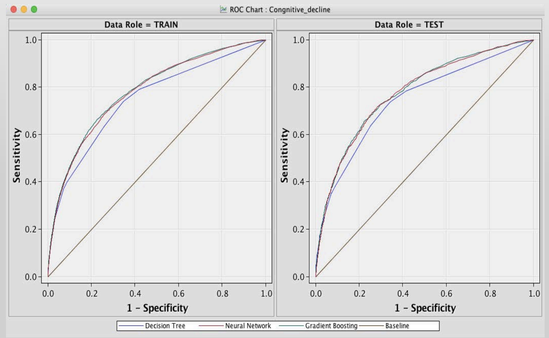


Fig 11: ROC Curve with imbalanced dataset

Feature importance generated from gradient boosting tree, the most important factors contribute to MCI are: Mental health, Married, Education level, Physical health, Exercise, Divorced, etc.

Conclusion

- SMOTEENN can significant improve recall which means enable to detect more people most likely to have MCI.
- All balanced strategies can improve recall value.
- Gradient boosting and neural networks are 2 best models to prediction MCI people.
- The top important features that can affect MCI are: Mental health, Marital status, Physical health, Exercise, etc.

Contact Information

Corresponding author: Meng Han: mhan9@kennesaw.edu
Liyuan Liu: lliyuan@students.kennesaw.edu
Yiyun Zhou: yzhou20@students.kennesaw.edu
Gita Taasoobshirazi: gtaasoob@kennesaw.edu

SAS® GLOBAL FORUM 2020

USERS PROGRAM

MARCH 29 – APRIL 1 | WASHINGTON, DC | #SASGF