CS699 Data Mining Project

Predicting Difficulty Living Independently

Prepared by

* Walter Stock
* Arfaan Jeddy

Institution

* Boston University, MET College

Course

* Data Mining
* CS699 A1, Spring 2025
* Instructor: Jae Young Lee

Date

* April 7th, 2025

Table of Contents

[Summary 2](#_Toc194779985)

[Datamining Tools 2](#_Toc194779986)

[Caret 2](#_Toc194779987)

[GgPlot2 2](#_Toc194779988)

[GridExtra 2](#_Toc194779989)

[ROSE 3](#_Toc194779990)

[RSample 3](#_Toc194779991)

[Split 3](#_Toc194779992)

[Tidyverse 3](#_Toc194779993)

[Algorithms 3](#_Toc194779994)

[Datamining Procedure 3](#_Toc194779995)

[1. Preprocessing – df\_preprocessed 3](#_Toc194779996)

[1. Preprocessing – 1. Removal of Columns without Value 3](#_Toc194779997)

[1. Preprocessing – 2. Data Types 3](#_Toc194779998)

[1. Preprocessing – 3. Data Dictionary 4](#_Toc194779999)

[1. Preprocessing – 4. Final Refinements 4](#_Toc194780000)

[1. Preprocessing – 5. Conclusion 4](#_Toc194780001)

[2. Split – df\_training, df\_testing 4](#_Toc194780002)

[3. Balance – df\_balanced1, df\_balanced2 4](#_Toc194780003)

[4. Select – df\_select#\_balanced# 4](#_Toc194780004)

[Datamining Results 4](#_Toc194780005)

# Summary

F

# Datamining Tools

## Caret

The Caret package is an R package for building, tuning, and evaluating machine learning models. This is the primary package used throughout the project.

## GgPlot2

The ggPlot2 package is an R package for generating plots. This package was used to generate visualizations for Fisher Scores and Correlations for attribute selection.

## GridExtra

F

## ROSE

## RSample

## Split

T

## Tidyverse

# Algorithms

## Attribute Selection Algorithms

### Missing Values

### Fisher

I chose the Fisher algorithm over the Chi-square algorithm because of its ability to handle sparse data.

### Correlation

# Datamining Procedure

This is a detailed overview of the procedure used to complete this project. The project was completed with a single R file, ***full\_project.R***, that is included with the submission. There are 5 sections: Preprocess, Split, Balance, Select, Model.

## Preprocessing – df\_preprocessed

Preprocessing includes loading the data provided for the project, ***project\_data.csv***, and then the following steps.

### Preprocessing – 1. Removal of Columns without Value

Columns without value are immediately removed to provide the greatest benefit. These columns all have the same value for all records and include: STATE, REGION, DIVISION, ADJINC, and RT.

### Preprocessing – 2. Data Types

Information about each column of data will be important throughout the project. Here we create a data frame, ***df\_columns\_info***, with every column name and the class of the column after import. All columns are reviewed and classified as Integer, Factor, Factor Levels, Logical.

### Preprocessing – 3. Data Dictionary

To assist with analysis and communication, the data dictionary for the data being analyzed is downloaded, ***PUMS\_Data\_Dictionary\_2023.csv,*** and loaded into data frame, ***data\_dict\_df***. This data frame is used to add new columns to the main data frame, ***df***, that include the description of each column and names of the values in that column (when available).

### Preprocessing – 4. Final Refinements

The class of all columns in the main data frame are updated based on the analysis completed and recorded in df\_columns\_info. Columns identified as likely to be good predictors are flagged as well. As a final step the df data frame is written to ***df\_preprocessed*** and exported as preprocessed\_data.csv and preprocessed.RData. The preprocessed.RData file will persist the data frame column classes.

### Preprocessing – 5. Conclusion

This completes preprocessing. It is important to note that the record count and original underlying values are unchanged. These will not be touched until after an unsupervised splitting of the data is completed in the next step.

## Split – df\_training, df\_testing

The data, ***df\_preprocessed***,is split into a training, ***df\_training***, and testing, ***df\_testing***, dataset using the Split function. The testing dataset will be used for performance evaluation. The Split function is an unsupervised method for splitting the data with equal weights from both classes, difficulty living independently and without difficulty. There is addition information about the Split function in the Datamining Tools section of this document.

## Balance – df\_balanced1, df\_balanced2

The data is made more balanced with respect to Class outcomes using down sample and up sample. SMOTE is not used due to the large amount categorical data in the dataset.

## Select – df\_select#\_balanced#

### Select 1 – Missing Value Removal

Selection methodology 1 goes through each column first removing any column with more than 5 percent of its values missing. It then goes through each row and removes rows with more than 1 percent of values missing. This is done for both balanced datasets.

### Select 2 – Fisher, Correlation, Outliers

Selection methodology 2 goes through 3 steps for each balanced dataset. Fisher and then Correlation and then analysis of distribution with outliers.

### Fisher and Correlation – Balanced Dataset 1

A graph with a number of data

AI-generated content may be incorrect.

# 

# Datamining Results