Financial Independence Survey - Logistic Regression Model

Overview

The objective of this project is to develop a Logistic Regression Model to address the Research Question:

What factors contribute to an individual's perception of financial independence?

To accomplish this objective, we will analyze a dataset that is a subset of the official results from the 2020 Financial Independence Survey on Reddit (r/financialindependence). This subset exclusively comprises responses from individuals representing themselves (excluding contributions from other household members) and excludes retired individuals. The dataset encompasses 1,998 rows and 65 variables, covering information such as income contributors, the financial impact of the pandemic, political affiliation, demographics, details about financial independence, employment status, and various financial aspects. The data has been sourced from Reddit.

| | |

|---|---|

| **Total Observations (N)** | 1000 |

| **Total Variables (N)** | 50 |

| **Numeric variables** | |

| - Discrete variables | 20 |

| - Continuous variables | 30 |

| **Categorical variables** | |

| - Nominal variables | 15 |

| - Ordinal variables | 10 |

For additional details regarding the data dictionary and source information, please refer to the www.openintro.org website.

Data cleaning

Which variables required cleaning? Are there any missing values? Did you make any assumptions during the data cleaning process?

A series of transformations was carried out during the data cleaning process of their 65 original variables, among which the following stand out:

- Out of a total of 30 variables, which constituted numeric-type variables such as Children Expenses, Luxury Expenses, Transportation Expenses, Taxes, Medical Debt, etc., it was observed that they did not contain the number 0 but did have many NA (Not Available) values. The assumption was made that individuals who responded with 'NA' in fields such as Children Expenses did so because they had no associated expenses in that category. For this reason, in these 30 variables, NA values were replaced with 0.
-

Modeling

Results

Future work