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. This subset 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.

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

Data cleaning

In the data cleaning process, the following transformations were implemented. For more details on the specific variables that experienced these transformations, please refer to the appendix page.

- The dependent variable **fin_indy** (Are you financially independent?) was transformed from a string variable to a binary numeric variable (0 and 1).
- The variable **political** had 24 categories, which were reduced to 4: "Democrat," "Libertarian Party," "Republican," and "Other." Finally, it was transformed from string to factor.
- The variable **age** was transformed from a string variable to a numeric variable.
- The variable **edu** had 11 categories, which were reduced to 4: "High School or Less", "Some College or Trade School", "Bachelor's or Associate's Degree", "Doctorate or Master's Degree". Finally, it was transformed from string to factor.
- The variable **housing** (What is your current housing situation?) had 19 categories, which were reduced to 3: "Live with family or friends", "Own", "Rent". Finally, it was transformed from string to factor.

- Out of a total of 29 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 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 29 variables, NA values were replaced with 0. The complete detail of the variables can be reviewed on the appendix page
- For the **cash** variable, the outliers that deviated more than 3 standard deviations from the mean were removed.
- The variable **total_assets** was created, which is formed by the sum of 8 variables such as cash, investment accounts, crypto, etc. or more information, refer to the appendix
- The variable **total_debts** was created, which is formed by the sum of 7 variables such as student loans, mortgage, medical debt, etc. or more information, refer to the appendix
- The variable **total_expenses** was created, which is formed by the sum of 12 variables such as necessities expenses, children expenses, transportation expenses, etc. or more information, refer to the appendix
- Finally, observations with NA values for any of the above variables were removed, resulting in 16 observations out of the original total of 1,998, leaving us with 1,982 final observations.

Modeling

To address our research question, "What factors contribute to an individual's perception of financial independence?" we will employ a Logistic Regression Model. This model is suitable for this situation as it allows us to examine how various predictor variables influence the probability of an individual perceiving financial independence. Given that the variable of interest is binary (yes/no), logistic regression will provide us with estimations of log probabilities and coefficients that will aid in understanding the direction and strength of the relationship between the considered factors and the perception of financial independence. This approach is particularly useful when seeking to identify key contributors to the perception of financial independence in a binary decision-making context.

luego de analizar las variables provistas en el dataset, cuales se podria considerar que pueden influenciar en la variable de independencia financiera, se seleccionaron a priori

Results

Future work

The analysis conducted provides us with a model that allows us to identify the factors influencing people's perception of financial independence. While the obtained results make sense, it's important to consider the limitations of our study.

One key consideration is the potential presence of confounding variables, such as age, which could impact an individual's levels of assets, liabilities, and income. These factors theoretically should influence financial independence. A future analysis could explore the results by removing the confounding variable of age to more precisely assess the impact of the specific factors considered in our model.

Furthermore, it would be relevant to conduct cross-validation analyses to evaluate the model's generalization capacity to unseen data. This helps verify if the model is robust and can be applied to new samples, which is crucial to ensure the reliability of conclusions and the practical utility of the model in different contexts.

In conclusion, by addressing these considerations and conducting additional analyses, we can enhance the robustness and applicability of our model to understand and predict the perception of financial independence.

Appendix

Variable	Transformation
political	Transformed into a factor.
	From the original 24 categories, it was grouped into 4:
	"Democrat"
	"Libertarian Party"
	"Republican"
	"Other"
age	It was transformed from string to numeric.
edu	Transformed into a factor.
	From the original 11 categories, it was grouped into 4:
	"High School or Less"
	"Some College or Trade School"
	"Bachelor's or Associate's Degree"
	"Doctorate or Master's Degree"
fin_indy	It was transformed from string to binary
housing	Transformed into a factor.
	From the original 19 categories, it was grouped into 3:
	"Live with family or friends"
	"Own"
	"Rent"
home_value	NA replaced with 0
	sum in total_assets variable
brokerage	NA replaced with 0
_accts_tax	sum in total_assets variable
retirement	NA replaced with 0
_accts_tax	sum in total_assets variable
cash	NA replaced with 0
	sum in total_assets variable
invst_accts	NA replaced with 0
	sum in total_assets variable
spec_crypto	NA replaced with 0
	sum in total_assets variable
invst_prop_bus_own	NA replaced with 0
	sum in total_assets variable
other_val	NA replaced with 0
	sum in total_assets variable
student_loans	NA replaced with 0
	sum in total_debts variable
mortgage	NA replaced with 0
	sum in total_debts variable

auto_loan	NA replaced with 0
	sum in total_debts variable
credit_personal_loan	NA replaced with 0
	sum in total_debts variable
medical_debt	NA replaced with 0
	sum in total_debts variable
invst_prop_bus_own_debt	NA replaced with 0
	sum in total_debts variable
other_debt	NA replaced with 0
	sum in total_debts variable
2020_gross_inc	NA replaced with 0
2020_housing_exp	NA replaced with 0
	sum in total_expenses variable
2020_utilities_exp	NA replaced with 0
	sum in total_expenses variable
2020_transp_exp	NA replaced with 0
	sum in total_expenses variable
2020_necessities_exp	NA replaced with 0
	sum in total_expenses variable
2020_lux_exp	NA replaced with 0
	sum in total_expenses variable
2020_child_exp	NA replaced with 0
	sum in total_expenses variable
2020_debt_repay	NA replaced with 0
	sum in total_expenses variable
2020_invst_save	NA replaced with 0
2020_charity	NA replaced with 0
	sum in total_expenses variable
2020_healthcare_exp	NA replaced with 0
	sum in total_expenses variable
2020_taxes	NA replaced with 0
	sum in total_expenses variable
2020_edu_exp	NA replaced with 0
	sum in total_expenses variable
2020_other_exp	NA replaced with 0
	sum in total_expenses variable