Financial Independence Analysis

Tina Yi

2023-10-18

Overview: The research uses the Reddit finance dataset, which is a Reddit survey on financial independence. This dataset comprised 1998 observations and 65 variables. The goal of this research is to investigate which factors contribute to whether or not someone considers themselves to be financially independent. This research investigates an inference problem.

Priori Variable Selection: From the original set of 65 variables, I refined the list to 45, grouping them into 12 unique categories. These categories covered a broad spectrum of factors influencing financial independence, such as the pandemic's effects, demographics, and financial status. I subsequently selected variables from this refined list, focusing on their relevance to financial independence in the Reddit Finance dataset, data integrity, non-redundancy, and potential confounding factors.

1. **Exclusion of Ambiguous Variables:** Variables difficult to interpret or categorize were excluded. For example, the 'age' variable was problematic due to its categorical nature with ten groups, including NAs. Converting it to a continuous variable could lead to misleading results. Uneven data distribution across age categories also made defining groups challenging. While I could group ages under and above 34 to minimize standard errors, there's no strong theoretical reason to choose 34 as a significant threshold in financial independence. Thus, I omitted the 'age' variable.
2. **Exclusion Based on Data Quality**: Variables with a significant number of missing values were left out to ensure the reliability of the dataset, such as 'Outstanding credit cards/personal loans' and 'Outstanding medical debt'. These variables, while potentially informative, carried the risk of introducing biases and limitations.
3. **Minimizing Redundancy:** For each category, a single, most representative variable was chosen. For example, 'pandemic financial impact' was chosen over similar variables like 'pandemic income change'. This approach was used across categories such as demographic, education, and financial to prevent overlapping measurements.
4. **Consideration of Potential Confounders**: Key variables known to be confounding factors were incorporated into the analysis. Variables like gender, race, and education, known to act as confounding factors, were included. For example, race/gender could confound the relationship between education and financial independence. These variables were incorporated to control for their effects and provide a comprehensive understanding of the factors influencing financial independence within the Reddit community.

I carefully selected an initial set of 11 predictor variables based on their expected relationship with financial independence.

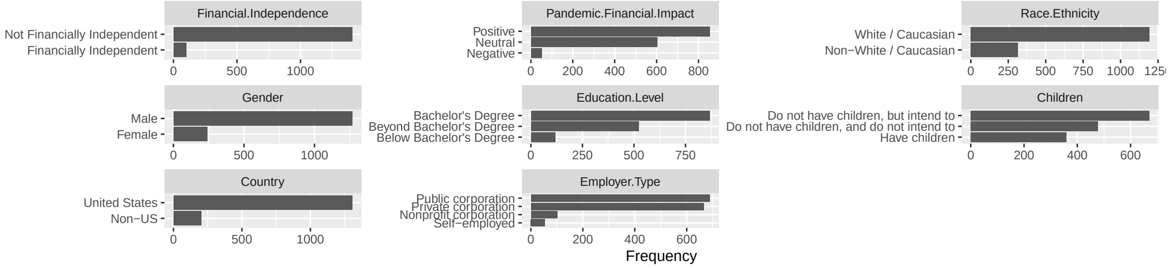
Binary.Outcome	Continuous.Predictors	Categorical.Predictors
Financial Independence	2020 Gross Income, Cash, 2020 Housing Expenditure, 2020 Necessities Expenditure	Pandemic Financial Impact, Race, Gender, Education, Children, Country, Employer

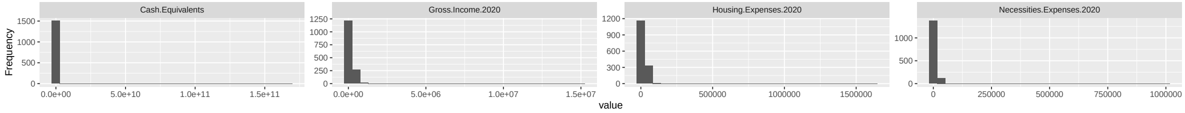
Data Cleaning: After undergoing the cleaning steps, the refined dataset now contains **12** variables with a total of **1,508** observations.

- Outcome variable: '**Financial Independence**' had no NAs and was categorized into "Financially Independent" (147 observations) and "Not Financially Independent" (1851 observations).

- Categorical predictor variables: (1) **Pandemic Financial Impact**: After filtering out 10 NAs, I grouped the data into three categories: “Positive” , “Negative” , and “Neutral”. (2) **Race**: I removed instances with 29 ‘N/A’, 17 ‘NA’, and 62 ‘Decline to State’. Given the significant disparity in the number of data points for “White/Caucasian” (1458 observations) compared to the other race categories, I made an executive decision to group the data into two broader categories: “White/Caucasian” and “Non-White/Caucasian”. This was based on both practical considerations of reducing standard error cause by small data size and an underlying assumption that financial independence perceptions might differ between white and non-white individuals. (3) **Gender**: After filtering out 2 NAs and 7 ‘Decline to State’, and due to concerns about small sample sizes of “Non-binary” group (9 observations) that lead to high standard errors, I decided to drop the “Non-binary” group and categorize gender into two main groups: “Male” and “Female”. (4) **Education**: After removing 3 NAs, I underwent an initial reorganization by merging similar educational categories. Further, I grouped these categories into “Bachelor’s Degree,” “Beyond Bachelor’s Degree,” and “Below Bachelor’s Degree” based on the assumption that having a bachelor’s degree may be a significant threshold affecting financial independence. (5) **Children**: I filtered out 43 ‘N/A’ and 7 ‘NA’ and categorized the data into three groups **“**Do not have children, but intend to” “Do not have children, and do not intend to” and “Have Children”. (6) **Country**: No NAs. I merged the “US Armed Forces” category with the “United States” category. I then grouped all remaining countries into a “Non-US” category. Thus, I ended up categorized the data into two groups “United States” and “Non-US”. This decision was driven by the situation that the US had 1554 observations, while most other countries had fewer than 10 observations each, potentially leading to large standard errors and bias. Additionally, this categorization was influenced by the assumption that individuals in the US may perceive themselves as more financially independent compared to those in non-US countries. (7) **Employer**: I removed 8 NAs and combined “public agency” into the “public corporation” category based on the assumption of both belonging to the public sector. Thus, I eneded up categorized the data into four groups “nonprofit corporation”, “public corporation”, “private corporation” and “self-employed”.

- Continuous predictor variables: Any missing values were dropped from the variables “**2022 Gross Income**” (NAs=38), “**Cash**” (NAs=35), “**2020 Housing Expenditure**”(NAs=87), and “**2020 Necessities Expenditure**”(NAs=117).





Modeling: Logistic regression is ideal for this research, aiming to identify factors influencing one’s perception of financial independence. The study has a binary outcome, perfectly suited for logistic regression, which predicts the likelihood of binary results using various predictor variables. It can highlight significant predictors and accommodate both continuous and categorical variables. Given the research’s binary nature, logistic regression is a fitting method for data analysis.

VIF: After initial model fitting, I employed GVIF to assess multicollinearity in the logistic regression. While most predictors showed minimal multicollinearity with GVIF close to 1, the variables 2020 gross income, 2020 housing expenditure, and 2020 necessities expenditure had slightly elevated GVIF values between 1.5-2, yet still within acceptable limits. All are below the threshold of concern.

Influential Points: After the initial modeling, I observed that children(Do not have children, but intend to) and 2020 housing expenditure have significant p-value(**), employer(Self-employed) have significant p-value(*), and 2020 gross income have significant p value (***). From the Cook’s distance plot, I checked if the logistic regression model have influential points that need to be handled. I observed that point 133 sticks out compared to the rest. I removed point 133 and reran the main regression. After removing point 133, I found that cash's p-value becomes significant (*) and 2020 housing expenditure's p-value significance decreased from (**) to (*). Therefore, point 133 is indeed influential. Then I removed point 155 that sticked out compared to the rest and reran the main regression. I found that cash's p value's significance increase from (*) to (***) and 2020 housing expenditure's p-value became insignificant after the removal of 155. This proves that 155 is indeed an influential point. Afterwards, I dropped point 635 that sticks out. Upon removing point 635 and observing my regression outcome, I concluded that point 635 is indeed influential as the p-value of 2020 housing expenditure and 2020 necessity expenditure becomes significant(*) after its removal, indicating that this point has a substantial impact on the model as a whole. I also removed point 608, and revisited the model. However, the significance level of the model remained unchanged, suggesting that point 608 is not an influential point as it doesn't have substantial impact on the model as a whole. I decided to keep point 608 and stop my influential points checking at this point. In the end, I removed three influential points, 133, 155 and 635.

Logistic Regression Assumptions: (1) **Linearity**: For this model, I assume a linear relationship between the log odds of the outcome and its predictors. Examining the predicted log odds plots revealed some outliers for continuous variables, yet they're not influential. Excluding these outliers, the data demonstrates good linearity. While outliers are common in financial variables, I chose not to transform predictors to preserve interpretive clarity. (2) **Independence**: Observations are independent of each other, which is addressed by the study design.

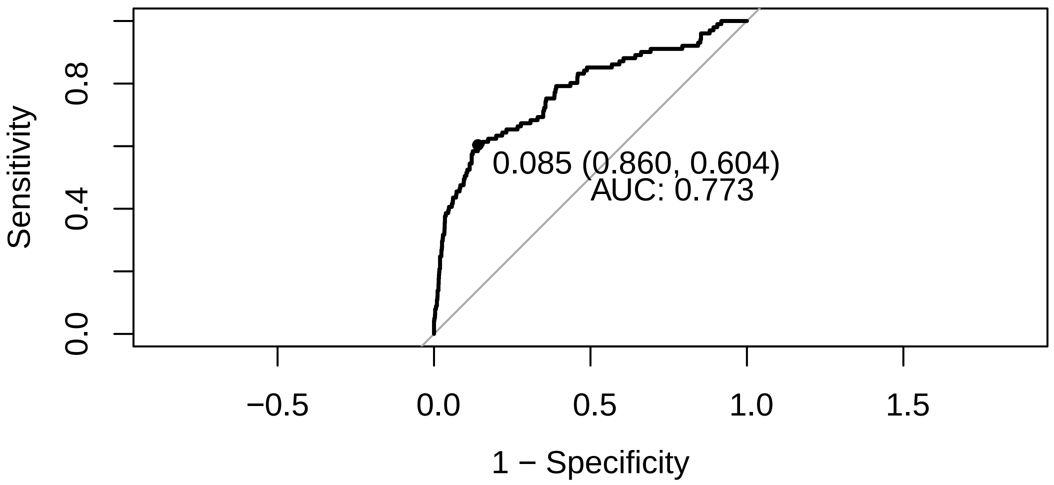
Summary Output Table:

	Financial Independence			
Predictors	Odds Ratios	std. Error	CI	p
(Intercept)	0.21	0.17	0.04 – 0.93	0.051
Pandemic Financial Impact [Neutral]	0.75	0.40	0.29 – 2.38	0.587
Pandemic Financial Impact [Positive]	0.67	0.35	0.26 – 2.11	0.445
Race [White / Caucasian]	0.66	0.17	0.40 – 1.11	0.104
Gender [Male]	0.93	0.28	0.53 – 1.72	0.807
Education [Below Bachelor's Degree]	1.01	0.43	0.42 – 2.21	0.974
Education [Beyond Bachelor's Degree]	1.16	0.28	0.72 – 1.86	0.535
Children [Do not have children, but intend to]	0.46	0.13	0.27 – 0.79	**0.005**
Children [Have children]	0.99	0.28	0.56 – 1.71	0.962
Country [United States]	0.74	0.24	0.40 – 1.44	0.350
Employer [Private corporation]	0.54	0.24	0.23 – 1.40	0.170
Employer [Public corporation]	0.96	0.42	0.44 – 2.43	0.924
Employer [Self-employed]	2.98	1.62	1.04 – 9.04	**0.045**
2020 Gross Income	1.00	0.00	1.00 – 1.00	**<0.001**
Cash	1.00	0.00	1.00 – 1.00	**<0.001**

2020 Housing Expenditure	1.00	0.00	1.00 – 1.00	**0.035**
2020 necessities exp	1.00	0.00	1.00 – 1.00	**0.042**
Observations	1505			
R^2 Tjur	0.127			

Model Assessment: To evaluate the performance of my model in determining financial independence, I used a combination of statistical metrics and visual tools, including ROC curve and confusion matrix.

- **Receiver Operating Characteristic (ROC):** For a visual assessment, I plotted a ROC curve utilized an optimal threshold of 0.085. The AUC stands at 0.773, suggests that while the model has some discriminatory power, there's room for improvement.



- **Confusion Matrix**: My primary tool was the confusion matrix, which provides a breakdown of where the model's predictions align with or deviate from the actual outcomes. In this confusion matrix, I used the best thresh hold 0.085.

Model Performance Metrics
'Positive' Class: Financially Independent

Metric	Value
Accuracy	0.8419
Kappa	0.2682
F1	0.33889

In the model, I aim to discern factors affecting perceptions of financial independence. With 101 samples labeled as "Financially Independent" versus 1404 as "Not Financially Independent", there's a clear class imbalance. Hence, I've emphasized overall accuracy and the F1 score, which adeptly handles class imbalances by balancing precision and recall. The kappa score further illuminates this skewness.

- **Accuracy**: 84.19% of the total predictions were correct. This suggests that the model is doing fairly well, but it's essential to look at other metrics as well, especially when classes are imbalanced.

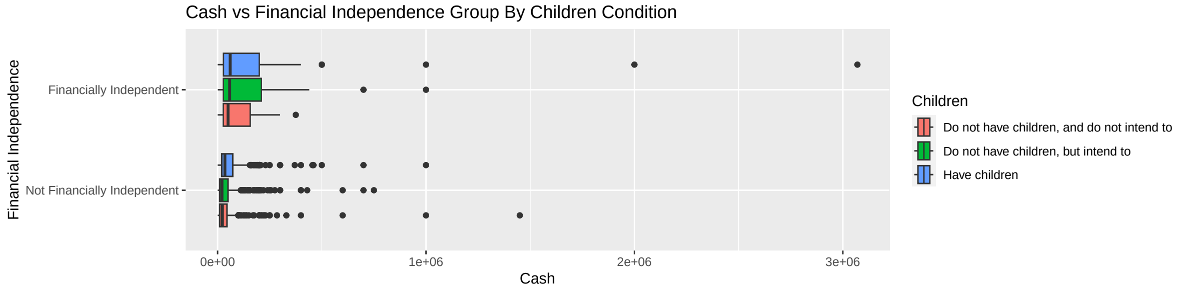
- **Kappa**: 0.2682 indicates a moderate agreement between actual and predicted classifications beyond chance. The Kappa score of 0.2682, though not high, suggests the model's predictions offer some advantage over random guessing, especially considering the imbalanced nature of the dataset.
- **F1 Score**: The harmonic mean of precision and recall is 33.89%. While not high, this score reflects the challenges of achieving a balance between precision and recall, especially in imbalanced datasets.

In summary, while the model has a decent accuracy and specificity, it faces challenges in striking a balance between identifying those who are "Financially Independent" and those who are not. This is evident in the relatively low precision and F1 score. The class imbalance further complicates the model's performance, emphasizing the need for a nuanced evaluation beyond mere accuracy.

Results: Based on the ouput table, it is evident that the variable children (don't have children, but intend to), employer (self-employed), 2020 gross income, cash, 2020 housing expenditure, and 2020 necessities expenditure significantly contribute to whether or not someone considers themselves to be financially independent, with p value each less than 0.05. The specific interpretation of these key results are listed below:

- The odds of being financially independent for people who do not have children, but intend to are 0.46 times the odds of being financially independent for people who do not have children, but do not intend to, all else held constant. (p value: 0.00485 **) (Confidence Interval: 0.26936385 to 0.7879359)
- The odds of being financially independent for people who are self-employed are 2.98 times the odds of being financially independent for people whose employer are non-profit corporation, all else held constant. (p value: 0.04524 *) (Confidence Interval: 1.04052411 to 9.0369451)
- With each additional 2020 gross income, the odds of being financially independent increase 1.0000028 times, all else held constant. (p value: 2.24e-05 ***) (Confidence Interval: 1.00000151 to 1.0000041)
- With each additional cash, the odds of being financially independent increase 1.0000039 times, all else held constant. (p value: 6.58e-06 ***) (Confidence Interval: 1.00000226 to 1.0000057)
- With each additional 2020 housing expenditure, the odds of being financially independent decrease 0.99 times, all else held constant. (p value: 0.03496 *) (Confidence Interval: 0.99996654 to 0.9999979)
- With each additional 2020 necessities expenditure, the odds of being financially independent decrease 0.99 times, all else held constant. (p value: 0.04193 *) (Confidence Interval: 0.99993695 to 0.9999903)

Visual Representation: The graph below shows that having more cash is associated with greater financial independence, which aligns with the regression analysis. While the box plot initially suggested that intending to have children is linked to higher financial independence compared to not intending to have children, the regression model, which considers multiple factors, indicates that intending to have children is associated with lower financial independence.



Conclusion/Future work: This analysis has both strengths and limitations. Its strengths include a robust variable selection from a sizable sample of 1505 and control for confounders. A significant limitation is the class imbalance: only 101 samples are 'Financially Independent' compared to 1404 that aren't. While the model is reasonably accurate, it struggles due to this imbalance. Future strategies could include resampling, using techniques like SMOTE, ensemble methods, or expanding the dataset to tackle the imbalance.