Analyzing the Impact of Financial Stability and Family Structure on Elder Health and Longevity: A Machine Learning Approach Using the RAND HRS Longitudinal File 2020

Yuzhen Zhou, Zeyuan Pang

Part 1. Executive Summary

The primary goal of our research is to analyze the impact of financial stability and family structure on the health outcomes. This investigation is crucial as nations globally grapple with the ramifications of an aging demographic, including heightened demands on healthcare systems, pension sustainability, and the overall welfare of the elderly population. Our study aims to provide evidence-based insights to guide policymaking in healthcare, social security, and family support initiatives, thereby facilitating more effective strategies to aid older adults.

For this analysis, we utilize the RAND HRS Longitudinal File 2020, which encompasses 15 waves of interview data collected over two decades. This comprehensive dataset is invaluable for research on health, family dynamics, retirement planning, employment history, and includes imputations for income, assets, and healthcare spending.

By examining the connections between economic status, family structure, and the health of the elderly, we aim to recommend targeted interventions that could improve life quality, reduce medical costs, and promote the sustainability of aging populations.

This project is of particular importance as it confronts a pressing challenge faced by East Asia: its rapidly aging population. This demographic transformation poses significant challenges for the social and economic progress of these countries, necessitating innovative approaches to ensure the well-being of the elderly and their families in a changing societal landscape.

Part 2. Data Exploration and Preprocessing

RAND HRS Longitudinal File 2020 is a huge dataset. It took weeks to read and understand the document and organize the data from different waves into a single dataset.

2.1 Data exploration

After reading the document, several variables of interest were selected for the initial analysis. The response variables are shown in Table 1:

VARIABLE CODE	CODE MEANING
SHLT	Self-rated Health Level
COGTOT	Cognitive Level
MSTOT	Mental Status Level

Table 1

The input variables are shown in Table 2:

VARIABLE CODE	CODE MEANING	VARIABLE CODE	CODE MEANING
BMI	Body Mass Index	**PRPCNT**	Number of Private Insurance Plans
INHPFN	Total Number of Helpers Ever Helped	***INHPE***	*Any employee of institution ever helped*
HHHRES	Number of People in Household	***HINPOV***	*Live in poverty*
HCHILD	Number of Children	***PENINC***	*Current receiving pension income*
LIVSIB	Number of Living Siblings	***HIGOV***	*Covered by government health insurance plan*
HAIRA	Individual Retirement Account Assets	***RETMON***	*Retirement Status*
HATOTB	Total Asset Amount	***SLFEMP***	*Self-Employment Status*
IEARN	Individual Income		
HITOT	Total Household Income		

Table 2

** Italic items on right side stand for binary categorical variables*

Table 3 in Appendix shows the statistical description of the features. Figure 1 and 2 in Appendix show the histogram of features.

According to the table and figures presented, it is observed that most ordinal predictive features exhibit a rightward skew. This trend is both normal and understandable when considering the distributions of earnings and assets. The imbalance distributions of categorical features also project the real society that a few people live in poverty, many elders are still working, and most are covered by government insurance plan.

2.2 Data Preprocessing

Group the data by categorical feature

The whole dataset is separated into 32 different groups according to the different combinations of the seven binary categorical features.

Outliers

Isolate Forest algorithm is applied to each group separately to detect and remove 10% of the total points as outliers. This method ensures a focused approach towards outlier detection and removal, allowing for a cleaner and more accurate analysis of the data within each group.

Isolation Forest is an efficient and specialized algorithm for anomaly detection, leveraging a tree-based approach that excels in identifying outliers with minimal assumptions about data distribution in high dimensional situation.

After removing the outliers, 22 groups of data with less than 500 samples are dropped to ensure the performance of machine learning and statistical accuracy of research result.

Feature selection

The column "INHPE" is dropped because its values are identical.

The summary of cleaned groups is shown below in Table 4

HINPOV	PENINC	HIGOV	RETMON	SLFEMP	COUNT
0	0	0	0	0	11859
0	0	1	0	0	4554
0	0	1	1	0	4067
0	0	1	1	1	2967
0	1	1	1	0	2723
0	0	1	0	1	2532
0	0	0	0	1	1969
0	1	1	1	1	1418
0	1	1	0	0	883
1	0	0	0	0	566

Table 4

Part 3. Model

3.1 Baseline Models

In this phase of the analysis, four baseline models were evaluated: the MLP (Multi-Layer Perceptron) regressor, the KNN (K-Nearest Neighbors) regressor, the Random Forest regressor, and the Linear regressor, the latter of which utilizes features generated through the application of the K-means algorithm. The cleaned dataset was divided into training and testing subsets following a 3:1 ratio, facilitating the training of the baseline models and the subsequent evaluation of their performance. The outcomes of these performance evaluations are detailed in Table 5.

Model	Test R-Squared	MSE
MLP regressor with default parameters	-49868.69	265301.27
KNN regressor with k = 2	0.220	4.92
Linear regressor & features generated by K-means with clusters = 50	0.073	5.79
Random Forest regressor with default parameters	0.656	2.18

Table 5

The Random Forest regressors exhibit superior performance compared to other baseline models, as evidenced by achieving the highest R-squared values. This algorithm's distinctive advantage lies in its robustness and efficiency in managing complex, unstructured data across multiple dimensions.

Figures 3, 4, and 5 in the Appendix, which present the Kernel Density Estimate (KDE) plots of errors for each data point during the Random Forest regressor fitting, further support this observation. The KDE plots approximate normal distributions, indicating that the Random Forest regressors have a commendable capability in accurately predicting the target feature.

3.2 Grouped Data

Upon deploying the Random Forest regressor on each cleaned dataset group, its efficacy is recorded in Table 5 of the Appendix. The division of data into test and training sets was managed through 5-fold cross-validation, ensuring a comprehensive evaluation by averaging the performance of these folds for a final measure. The test R-Squared values, predominantly exceeding 0.6 across the groups, suggest the model achieves a commendably high predictive accuracy. This demonstrates the Random Forest regressor's capability to effectively capture the underlying patterns within the dataset, as evidenced by the robust R-Squared metrics.

However, the noticeable disparity between the test and train performance within each group suggests the model is overfitting. This overfitting signifies that while the model predicts the training data well, its ability to generalize to unseen data is compromised, highlighting the need for further adjustments to improve its generalization capabilities.

3.3 Machine Learning Morphisms

ML1

ML1 is the unsupervised Isolation Forest Algorithm that maps each row of dataset to 0 or 1. It do not need any estimation on prior distribution, and it has no loss function. Isolation Forest use isolation score to decided whether a point is outlier.

$$ML_1 = (X \in \mathbb{R}^{18}, Y \in \{0,1\}, F(x, \mathit{\Theta}) = \text{Isolation Score Function}, P_\theta(\theta) = 1, L = \text{None})$$

ML2

ML2 is the Random Forest Regressor. It takes predictive features and predict response features by the average of every tree's output. It does not require a prior distribution but do need MSE as loss function to optimize the result.

$$ML_2 = \left(X \in \mathbb{R}^{15}, Y \in \mathbb{R}^3, F(x, \Theta) = \frac{1}{k} \sum_{i=1}^{k} T_i(x), P_\theta(\theta) = 1, L = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - F(x_i, \Theta)\big)^2 \right)$$

Part 4. Next Steps

4.1 More data preprocessing

More techniques that are used before fitting model would be considered, like standardization or other outlier detection methods.

4.2 Hyperparameter tunning

The model shows signs of overfitting. To counter this, we plan to first simplify the model to reduce its complexity, possibly through feature reduction or regularization. Following this, we'll implement grid search coupled with cross-validation to find the best hyperparameters. This two-step approach aims to balance the model, improving its performance on unseen data while preventing overfitting.

4.3 Relationships between predictive and responsive features

The feature importance metrics generated by a Random Forest regressor offer valuable insights into the relative significance of each feature in predicting the target variable, quantified as ratios. However, these ratios alone do not distinguish between the directionality of the relationships—that is, whether a feature positively or negatively impacts the target feature. To derive deeper insights from the data, it becomes necessary to explore beyond mere importance ratios. Techniques such as partial dependence plots (PDPs) or correlation analysis could be employed to discern the nature of the relationship between predictive features and the target. This additional layer of analysis will enable us to not only understand which features are important but also how they influence the target feature, providing a more nuanced understanding of the dataset and facilitating the development of more informed strategies and decisions.

4.4 Diagnose and prescriptive analysis

We plan to delve into the reasons behind the specific characteristics and distributions of the data and weights observed within each group. This investigation aims to uncover underlying patterns or issues that may inform our recommendations for policies or business strategies. Additionally, by conducting a comparative analysis of the differences across groups, we intend to deepen our understanding of the dataset. This enhanced insight will not only shed light on the unique attributes of each group but also potentially reveal broader trends or anomalies that could inform more targeted and effective interventions.

Part 5. Source Code

https://github.com/YuzhenZhou1327/ESE527_Project_HRS

Part 6. Appendix

Table 3

	count	mean	std	min	25%	50%	75%	max
SHLT	38487	2.475251	0.970384	1	2	2	3	5
BMI	38487	28.25911	5.320587	9.7	24.6	27.4	31.1	62.3
MSTOT	38487	13.36553	1.874137	2	12	14	15	15
COGTOT	38487	23.94676	4.143787	3	21	24	27	35
INHPFN	38487	0.041287	0.255348	0	0	0	0	5
INHPE	38487	2.60E-05	0.005097	0	0	0	0	1
HHHRES	38487	2.678879	1.140705	2	2	2	3	14
HCHILD	38487	3.26661	1.933677	0	2	3	4	11
LIVSIB	38487	2.944813	2.451244	0	1	2	4	18
HINPOV	38487	0.035518	0.185089	0	0	0	0	1
HINPOVA	38487	0.035544	0.185154	0	0	0	0	1
HAIRA	38487	78742.64	283976.1	0	0	0	60000	35027000
HATOTB	38487	579882.2	1330807	-3624527	76000	228400	588500	42226312
IEARN	38487	31068.2	52357.43	0	0	15000	42000	2000000
HITOT	38487	102512.5	159141.1	0	41812	70880	119400	10036000
PENINC	38487	0.167953	0.373829	0	0	0	0	1
HIGOV	38487	0.561618	0.496195	0	0	1	1	1
PRPCNT	38487	0.786214	0.620732	0	0	1	1	23
SLFEMP	38487	0.280484	0.449242	0	0	0	1	1
RETMON	38487	0.343285	0.474812	0	0	0	1	1

Figure 1 Histogram of Numerical Ordinal Features

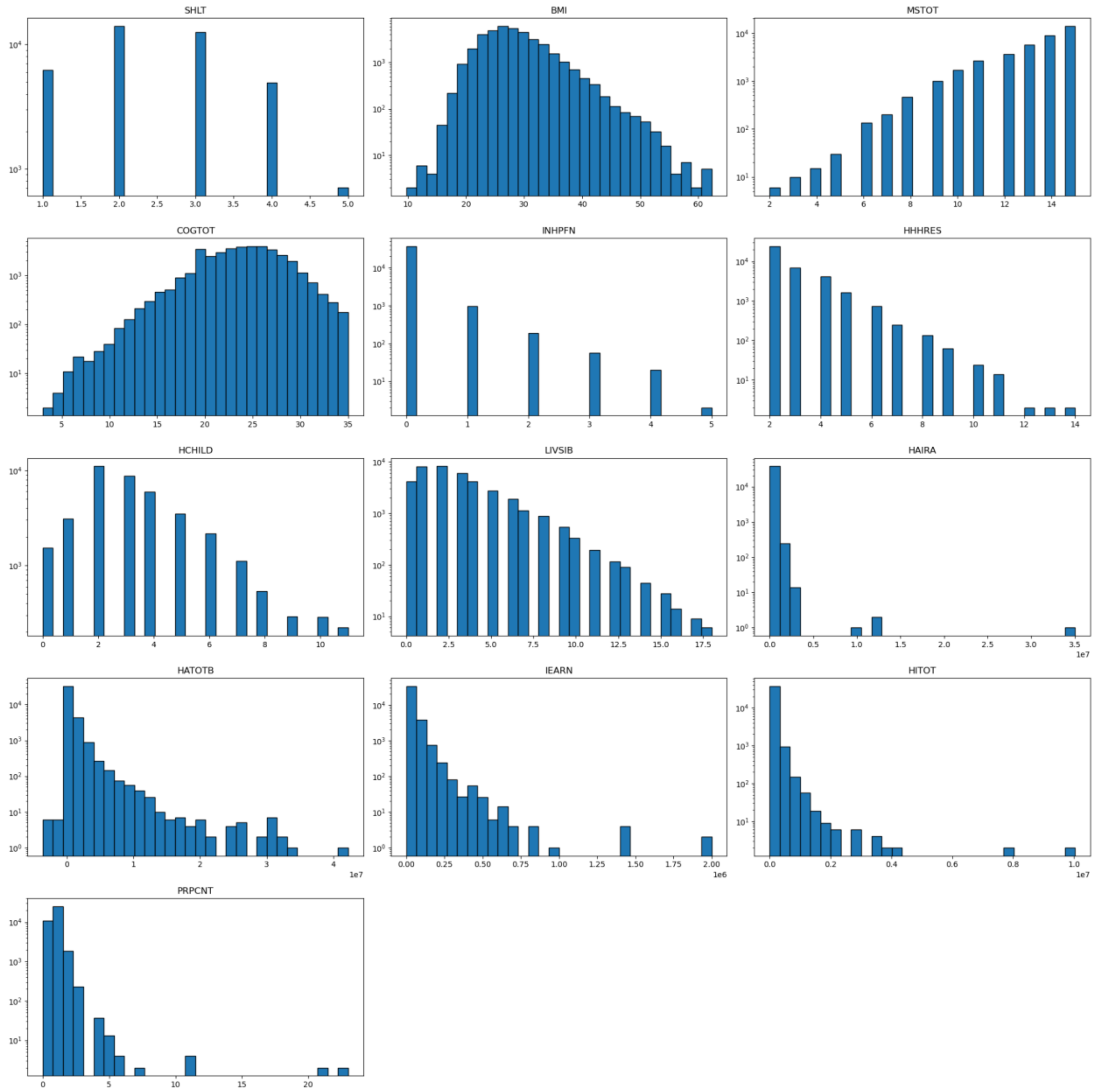

Figure 2 Histogram of Binary Categorical Features

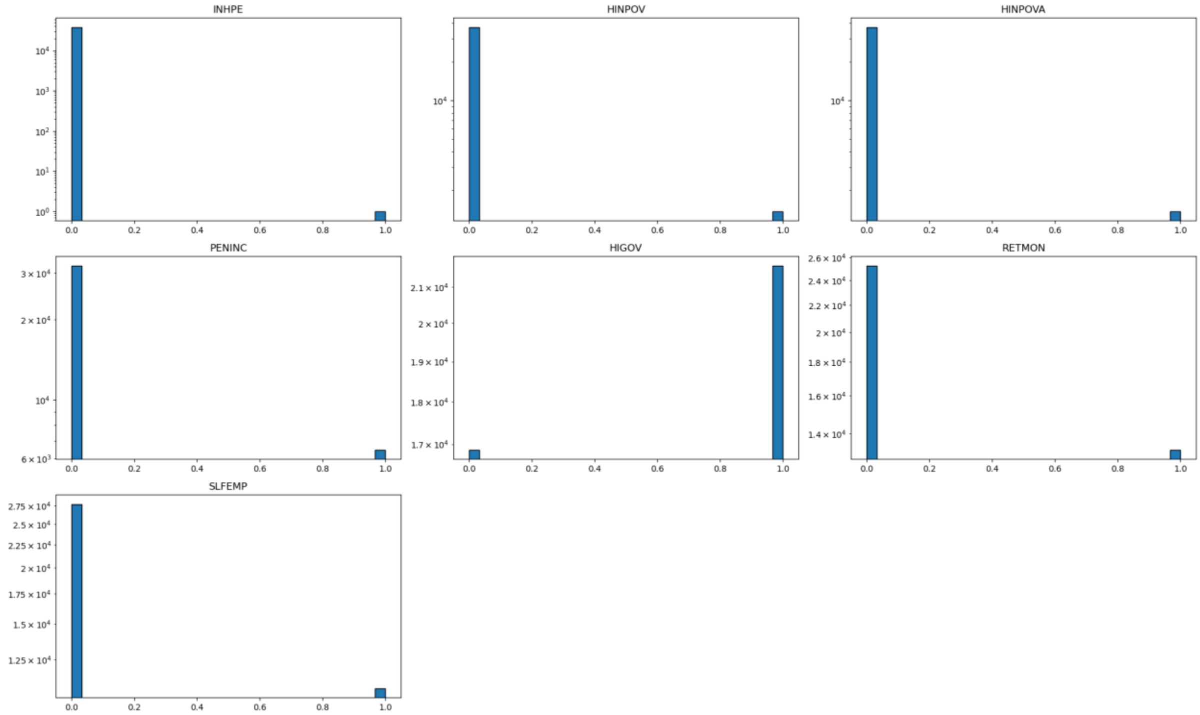

Figure 3

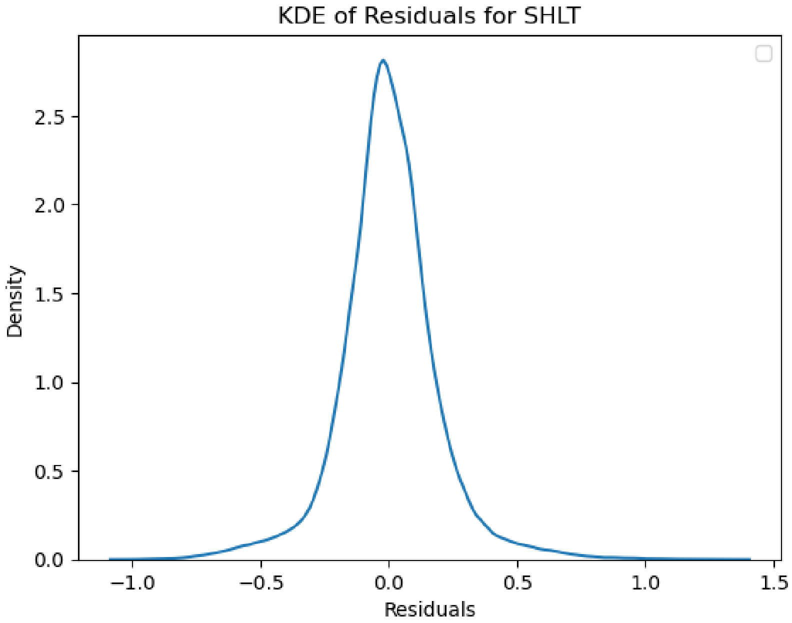

Figure 4

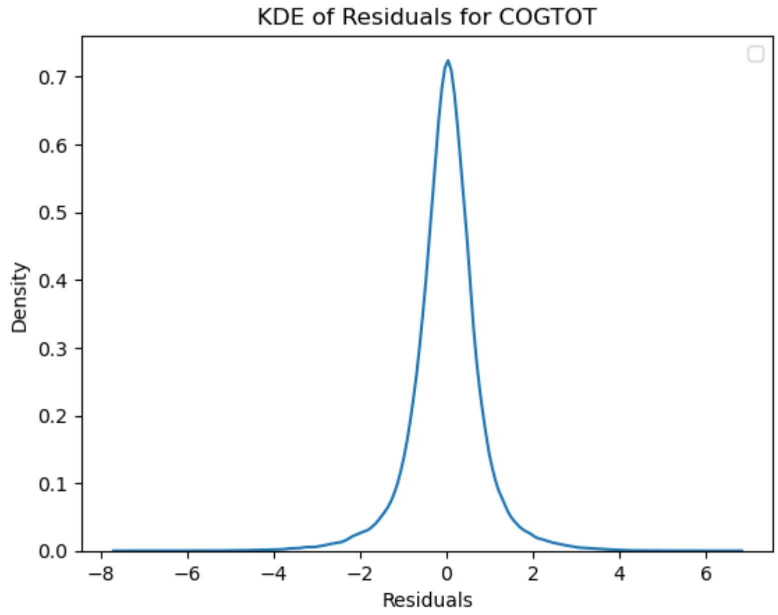

Figure 5

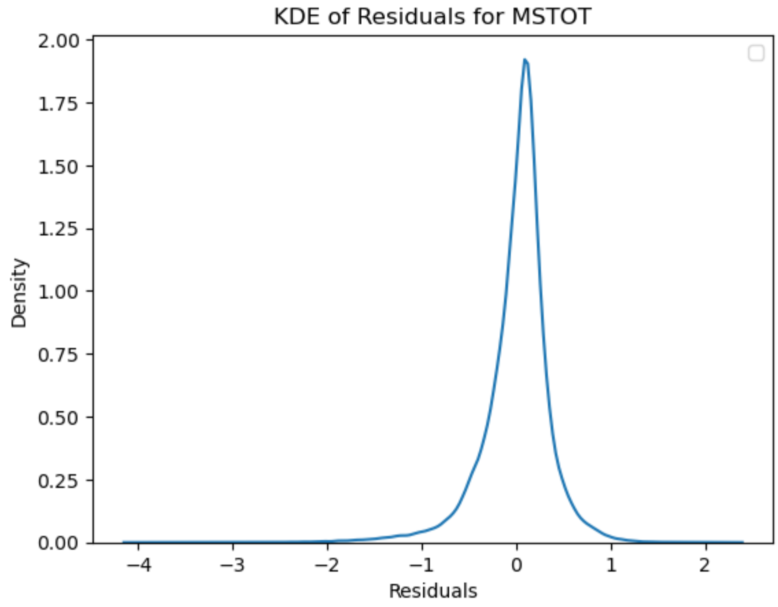

Table 5

Group	Train R-Squared	Train MSE	Test R-Squared	Test MSE
(0.0, 0.0, 0.0, 0.0, 0.0)	0.956115	0.2975	0.686855	2.1516
(0.0, 0.0, 0.0, 0.0, 1.0)	0.952207	0.308819	0.646954	2.287411
(0.0, 0.0, 1.0, 0.0, 0.0)	0.956238	0.341657	0.706917	2.32757
(0.0, 0.0, 1.0, 0.0, 1.0)	0.957859	0.331887	0.673744	2.575684
(0.0, 0.0, 1.0, 1.0, 0.0)	0.956421	0.301337	0.689104	2.166473
(0.0, 0.0, 1.0, 1.0, 1.0)	0.954554	0.32576	0.645943	2.581823
(0.0, 1.0, 1.0, 0.0, 0.0)	0.951703	0.338242	0.644797	2.4502
(0.0, 1.0, 1.0, 1.0, 0.0)	0.952756	0.284202	0.702184	1.758168
(0.0, 1.0, 1.0, 1.0, 1.0)	0.950648	0.301712	0.659185	2.121543
(1.0, 0.0, 0.0, 0.0, 0.0)	0.925576	0.635172	0.520246	4.057652