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

Authors : Yuzhen Zhou, Zeyuan Pang

School : McKelvey School of Engineering

Department : Electrical and System Engineering

Major : Engineering Data Analytics and Statistics

Advisor : Patricio S. La Rosa

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

Grouping Data by Categorical Features

The whole dataset is separated into 32 different groups according to the different combinations of the seven binary categorical features. The five categorical features are then replaced by a single group feature.

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. 10 groups remained.

Selecting and Merging Groups

The column “INHPE” is dropped because its values are identical.

To ensure the selected groups provide meaningful insights, a Multivariate Analysis of Variance (MANOVA) was conducted across each pair of groups. The findings are presented in Table 4. Adopting a significance level of 0.1, the analysis revealed that the three pairs of groups without statistically significant differences are:

(0,0,1,0,1), (0,1,1,0,0) with p-value of 0.103
(0,0,1,0,1), (0,1,1,1,0) with p-value of 0.109
(0,1,1,0,0), (0,1,1,1,0) with p-value of 0.648

Subsequently, these groups were amalgamated. There are 8 groups in total.

The summary of cleaned groups after group merging is shown below in Table 5

HINPOV	PENINC	HIGOV	RETMON	SLFEMP	COUNT
0	0	0	0	0	11859
0	0	1	0	1	
0	1	1	0	0	6138
0	1	1	1	0	
0	0	1	0	0	4554
0	0	1	1	0	4067
0	0	1	1	1	2967
0	0	0	0	1	1969
0	1	1	1	1	1418
1	0	0	0	0	566

Table 5

** The groups in the box are combined*

To examine the interrelationship between each pair of target features, correlation matrices for the three designated target features were constructed for each group. Additionally, the Pearson correlation coefficient, along with its corresponding p-value for each pair of target features within each group, was calculated. The findings from these analyses are presented in Figures 3 and 4 of the Appendix. These results reveal that MSTOT and COGTOT exhibit a medium to strong correlation across the eight groups. However, the accompanying p-values are exceedingly small, denoting a statistically significant difference between MSTOT and COGTOT across all groups. Consequently, based on this statistical significance, the decision was made not to amalgamate the three target features.

Feature Transformation

Upon completion of the model fitting, the presence of heteroskedasticity in the predictions was identified. To address this issue, a logarithmic transformation was implemented. This transformation necessitates that all data be non-negative. In cases where data points are negative but close to zero, it is feasible to adjust these values by adding a constant, allowing them to meet the non-negativity requirement for logarithmic transformation.

However, the feature “HATOTB” presents a challenge, containing approximately 1600 entries of negative data with substantially large absolute values, which complicates straightforward adjustment. Consequently, all samples, including negative values are retained but restricts the logarithmic transformation to features excluding “HATOTB”.

Part 3. Modeling Approaches

3.1 Descriptive methods

In the data exploration phase, a histogram is employed to investigate the distribution of individual variables. To identify the similarities across groups and variables, both Multivariate Analysis of Variance (MANOVA) and Pearson Correlation Test are utilized.

Additionally, descriptive statistical measures, such as the mean, standard deviation, and quartiles, are computed to illustrate the data's distribution. This approach is foundational in ensuring a comprehensive understanding of the dataset's characteristics.

3.2 Model Selection

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 7.

Model	Test R-Squared	MSE
MLP regressor with hidden layer structure (100,100,100,100) and relu activation	-28.02	92.72
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 7

MLP is a type of neural network known for its ability to model complex non-linear relationships within data. It is selected as a baseline model because it can capture intricate patterns through its layers and neurons. However, an R-squared of -28.02 indicates that the MLP regressor performed worse than a simple mean-based model, which might suggest that this problem is not well-suited for MLP. Also, it takes much longer time to train MLP than other models.

KNN is a non-parametric model that is simple to implement and understand. It makes predictions based on the proximity of data points, which can be effective for datasets with meaningful distance metrics. An R-squared of 0.220 shows that the KNN has limited predictive power in this setting but still provides some degree of accuracy.

Combining a linear regressor with feature engineering through K-means clustering can reveal underlying patterns within the data. The linear model provides a simple and interpretable baseline that assesses whether relationships in the data are linear. The R-squared of 0.073

indicates that this combination captures a slight portion of variance within the data but is not very effective.

Random Forest Regressor is a robust ensemble method that can handle non-linear relationships and interactions between variables. It is less prone to overfitting due to its ensemble nature and is often used as a strong baseline in predictive modeling. An R-squared of 0.656 indicates that the Random Forest model performed relatively well, capturing a significant portion of the variance in the dataset. The low MSE suggests that the predictions were close to the actual values on average.

Therefore, Random Forest Regressor is selected as the model for prediction on this problem since it is the best performing model among the ones selected, suggesting that the data has a structure that benefits from the ensemble approach this model uses.

3.3 Training and Predicting Target Features

Random Forest Regressor is trained separately for each group on the preprocessed dataset. For each group, the data is separated into training data and testing data with the ratio 9:1, and a grid search with 10-fold cross validation was performed on the training data. After finding the best parameters, the performance of Random Forest Regressor is evaluated based on testing data.

Three metrics are used to evaluate the performance—R-squared, RMSE, and MAPE. R-squared is best for understanding the proportion of variance explained by the model, RMSE is valuable for capturing the average error magnitude while penalizing large errors, and MAPE is useful for comparing the accuracy of models in terms of percentage errors, making it intuitive for expressing how large the errors are relative to actual values.

Tables 8 in the Appendix present the optimal parameters for each group, along with the corresponding performance metrics. The tables indicate that, overall, the Random Forest Regressor achieves superior performance. The model achieved an average test R-squared value of 0.73 and an average test Mean Absolute Percentage Errorof 0.011. Furthermore, the model demonstrated a robust ability to generalize to unseen data, as evidenced by the test R-squared being 20% higher than the training R-squared on most groups. This disparity suggests that the model not only fits the training data well but also effectively predicts outcomes in new, unexplored datasets, highlighting its utility in practical applications.

3.4 Prescriptive Methods

The feature importance derived from the Random Forest Regressor provides a quantitative analysis of the extent to which each input feature influences the target variables. This metric is crucial for understanding the predictive dynamics within the model, as it helps identify the most significant drivers of outcomes. The feature importance is shown in Table 10 of Appendix.

Additionally, by combining this with scatter plots that display the relationships between input features and the predicted or actual targets, researchers can explore the trends and directional influences of these features. Such visual representations enable a more nuanced investigation into how changes in input variables correlate with shifts in output, thereby offering valuable

insights into the behavior of the model under various conditions. This methodical approach enhances the interpretability of the model and supports more informed decision-making in feature selection and model refinement. An example scatter plot is shown below in Figure 7. All the scatter plots could be viewed on GitHub through here (link).

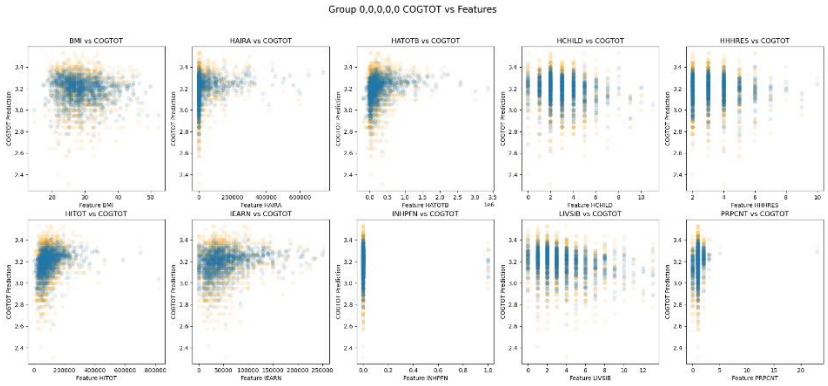

Figure 7

3.5 Machine Learning Morphisms

ML1

ML1 is the grouping process, generating a column of group label with the five binary categorical features.

$ML_1 = (X \in \{0,1\}^5, Y \in \text{Label}, F(x,\Theta) = \text{None}, P_\theta(\theta) = 1, L = \text{None})$

ML2

ML2 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 decide whether a point is outlier.

$ML_2 = (X \in \mathbb{R}^{18}, Y \in \{0,1\}, F(x,\Theta) = \text{Isolation Score Function}, P_\theta(\theta) = 1, L = \text{None})$

ML3

ML3 is the log transformation on continuous features.

$ML_3 = (X \in \mathbb{R}^{10}, Y \in \mathbb{R}^{10}, F(x,\Theta) = \log(x+10), P_\theta(\theta) = 1, L = \text{None})$

ML4

ML4 is the Random Forest Regressor. It takes predictive features and predict response features by the average of every tree output. It does not require a prior distribution but do need MSE as loss function to optimize the result.

$$ML_4 = \left(X \in \mathbb{R}^{10}, Y \in \mathbb{R}^3, F(x,\Theta) = \frac{1}{k}\sum_{i=1}^{k} T_i(x), P_\theta(\theta) = 1, L = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - F(x_i,\Theta)\right)^2\right)$$

Part 4 Result and Insights

4.1 Potential Population in Each Group

According to Table 2 and 5:

“0,0,0,0,0” could be the population of full-time employees of private companies or business who receive benefits like health insurance through their employer rather than the government.

“0,0,1,0,0” could be old government employees who receive health insurance through their employer or non-retired individuals who qualify for government health insurance due to other factors.

“0,0,1,1,0” could be a retiree who does not receive pension income but may have other sources of funds or retirement savings. They are covered by a government health insurance plan, likely due to their retiree status. People in this group might be enjoying their retirement with sufficient savings and investments to not require pension income, or they might be receiving Social Security benefits, which are not categorized as pension.

“0,0,1,1,1” could be elders retired from their primary career but have chosen to start their own business or continue working in a self-employed capacity. This is common for retirees who may want to stay active, pursue a passion, or supplement their income. The lack of pension income might mean they are relying on savings, investments, Social Security, or the income from their self-employment to fund their retirement.

“0,0,0,0,1” could stand for old people who are entrepreneurs or freelance professionals. Since they are not retired and don’t receive pension income, their primary source of income would be from their business activities.

“0,1,1,1,1” could be a retiree who, while receiving pension income and being officially retired, has chosen to continue working or to start working again in a self-employed capacity. This work could be in a field related to their previous employment or could be a new venture entirely. The government health insurance plan coverage is in line with their retired status.

“1,0,0,0,0” could be the old people who are employed in low-income jobs and do not have sufficient resources to be above the poverty line. They also could be unemployed and living in poverty without access to government-provided health insurance. This combination reflects a vulnerable segment of the population that could be the focus of targeted social support programs or policy interventions aimed at improving their economic status and access to health care.

The combined group “0,0,1,0,1+0,1,1,0,0+0,1,1,1,0” could represent a segment of the population that is somewhat transitionary, with some members still participating in the workforce and others who have retired. The inclusion of both pension-receiving and non-pension-receiving individuals suggests a diversity in terms of financial status and planning. It is a group that, overall, has health care needs covered by the government, which is a common characteristic in many countries for those who are older, regardless of employment status.

4.2 Insights from Feature Importance

According to Table 10 and scatter plots:

“BMI” is consistently one of the most important features across all groups, suggesting a significant influence on the target features. This indicates that health or physical conditions, as approximated by “BMI”, might have a strong predictive value for the targets. However, the distributions of “BMI” versus target features are almost Gaussian with no obvious linear relationship, suggesting that this relationship is complex and possibly influenced by multiple interacting factors. Addressing these complexities in analysis and policy could lead to more effective health interventions and better understanding of health dynamics in populations.

“HAIRA”, “HITOT”, and “HATOTB” appear prominently in most groups, particularly in those that include individuals with government health insurance (“HIGOV” = 1). This could suggest that benefits, possibly related to healthcare or personal care, play a crucial role in the well-being or conditions represented by the target features. Also, the scatter plots suggest that a higher value of financial-stability-related features is usually accompanied by a higher lower bound of Cognitive Level (“COGTOT”) and Mental Status Level (“MSTOT”) across all the groups, implying that programs designed to increase financial literacy, asset building, and economic opportunities could indirectly contribute to better cognitive and mental health outcomes, especially in lower-income communities.

In all groups, a higher individual income (“IEARN”) corresponds to an increased lower bound of mental status level (“MSTOT”) and a decreased upper bound of self-evaluated health level (“SHLT”). This observation agrees with the previous statement that greater individual income may provide better mental health outcomes. This could be due to reduced financial stress, better access to mental health resources, or a higher capacity to manage life's challenges effectively. The relationship where higher “IEARN” relates to a lower upper bound of “SHLT” is intriguing. It might suggest that while individuals with higher earnings evaluate their health more critically, or they may be more aware of health issues due to better access to health care and thus report more health problems. Alternatively, higher-income individuals might have jobs that are more demanding, leading to increased stress and perceived worse health despite objectively better health metrics.

However, the importance of ‘IEARN’ is relatively low among self-employed individuals (“SLFEMP”=1), while the importance of “HATOTB” and “HITIT” is high. It may because individual earnings may not capture the full economic picture as effectively as total household income or asset amount for self-employed individuals. This could be due to variability in self-reported earnings, income stability, or the way earnings are structured in self-employment. These observations suggest that the role of economic stability and variability in income should be considered when designing health policies or workplace wellness programs. For self-employed individuals, policies might need to focus more on providing stability and access to health resources.

The number of alive siblings (“LIVSIB”) shows a medium importance for each group, especially for “1,0,0,0,0” where “LIVSIB” is the fourth important feature. However, the scatterplots between “LIVSIB” and target features show a uniform distribution, without any

linear relationship. The influence of “LIVSIB” might be non-linear, meaning that there could be thresholds or complex interactions with other variables that affect the health outcomes.

The low importance scores for “INHPFN” suggest that this feature have a minimal impact on the prediction of the target variables due to its limited variability. It suggests that the impact of volunteer helpers is not captured adequately by the data or that their impact is minimal, and this feature is redundant in this problem.

The significance of the “PRPCNT” variable is generally low. However, an observable correlation exists between the number of private insurance plans and the baseline mental status level across various groups, except for the group “1,0,0,0,0”. Scatterplot analyses indicate that an increase in the number of private insurance plans is associated with an elevation in the lower boundary of mental status level. There may be a need to specifically target mental health support programs to individuals who have fewer private insurance plans, especially in economically disadvantaged groups. These programs could focus on providing affordable mental health services and reducing the financial burden of healthcare. Also, for poor groups like 1,0,0,0,0, interventions may need to be more holistic, addressing not just health insurance coverage but also employment opportunities, income support, and other social services to improve overall well-being and mental health.

4.3 Notable In-group and Inter-group Insights

In group “0,0,1,1,1”, fewer than 2 children or exceeding 5 children in the household are associated with an elevated lower bound of mental status level. Additionally, surpassing 5 children in the household correlates with an increased lower bound of cognitive level. It suggests that both ends of the family size spectrum might be providing environments conducive to better mental health. Fewer children may mean fewer caretaking responsibilities and more resources per child, potentially reducing stress and allowing more focus on personal mental well-being. Conversely, a very large family might foster a supportive environment that can be mentally stimulating and emotionally fulfilling, contributing positively to cognitive and mental health.

For retirees, having a larger family may provide not only emotional but also practical support, which could be particularly beneficial if they lack pension income and rely on other sources such as savings or self-employment. The interactions within a large family might help maintain cognitive functions and mental health due to increased social interactions and responsibilities. Hence, community and social programs that mimic the benefits of a large family could be beneficial. Programs that increase social interactions, provide community support, and engage them in mentally stimulating activities could replicate these benefits. Also, policies aimed at supporting retirees in maintaining an active lifestyle, whether through community involvement or family engagement, could be crucial. Incentives for multi-generational living or community-based family interactions might also be beneficial, especially for those who are self-employed and may not have as many opportunities for social interaction outside the home.

The relationship between “MSTOT” and “HHHRES” on most groups, including “0,0,1,1,0” and the combined group “0,0,1,0,1+0,1,1,0,0+0,1,1,1,0”, strengthens the statement above. For

these groups, more people living in a household relates to a higher lower bound of mental status level. It could also because that individual stressors related to daily living or financial burdens may be diluted as responsibilities and resources are shared among more members with more people in the household. Overall, it emphasizes the importance of social support systems and community programs that foster interaction and engagement among community members.

Part 5 Conclusions

5.1 Project Achievement

This project has thoroughly investigated the impact of financial stability and family structure on the health outcomes of the elderly using the RAND HRS Longitudinal File 2020. Through rigorous data preprocessing, model evaluation, and analysis, we have gained valuable insights into how economic and familial factors influence elderly health, which can inform targeted interventions and policy measures.

The research findings highlight the strong influence of financial assets and income on health, suggesting that enhancing economic stability could improve health outcomes for the elderly. For instance, higher total household income and asset amounts correlate with better cognitive and mental health statuses. These insights could guide the development of social security measures and financial assistance programs tailored to the needs of the elderly population, potentially leading to significant reductions in healthcare costs and improvements in life quality.

Furthermore, the analysis has uncovered the nuanced roles that family structure plays in the well-being of the elderly. For example, living arrangements and the number of children have been shown to impact mental health positively, providing a case for promoting community and family-based support systems.

5.2 Challenges and Difficulties

Addressing the sheer magnitude of the original dataset proved time-consuming when extracting pertinent and valuable information. Thoughtful deliberation was necessary for feature selection and the meticulous cleaning of empty or erroneous values.

Furthermore, certain features exhibited skewed distributions, leading to pronounced heteroskedasticity in the residuals of our model's predictions, consequently impacting its accuracy. Following consultation with our advisor, this issue was successfully mitigated through the implementation of log transformations on the features.

5.3 Future Possible Improvement

The process of model selection can be enhanced by optimizing each model to its peak performance before making a selection. Additionally, the stages of feature selection and hyperparameter tuning can benefit from the implementation of nested cross-validation. This approach not only enhances the accuracy and validity of the performance assessments but also boosts the generalizability of the models. This method ensures that the evaluation of the model's performance is not overly optimistic and is indicative of its capability to generalize to new, unseen data.

Part 6. Source Code

https://github.com/YuzhenZhou1327/ESE527_Project_HRS

Part 7. Appendix

Figure 1 Histogram of Numerical Ordinal Features

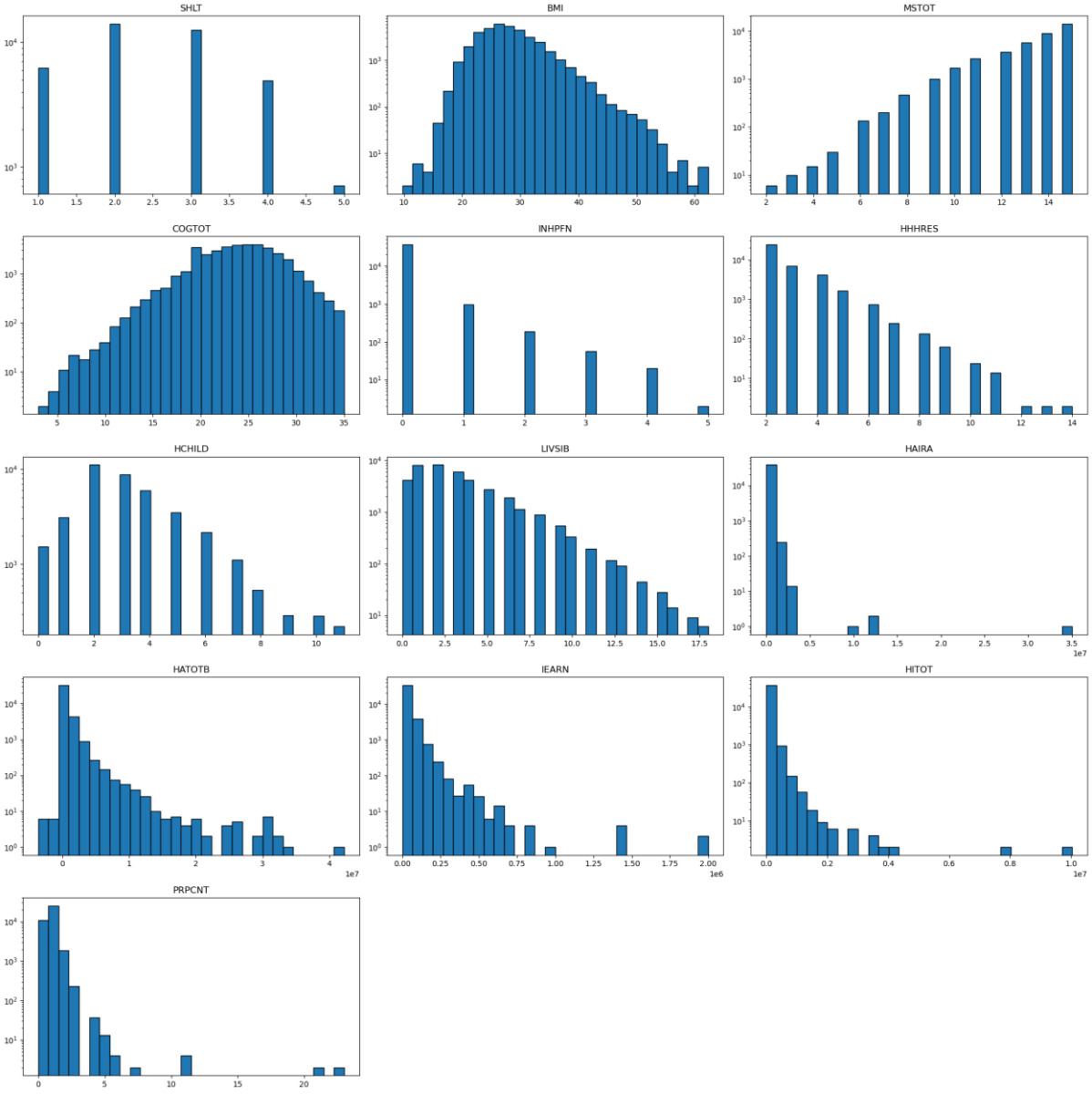

Figure 2 Histogram of Binary Categorical Features

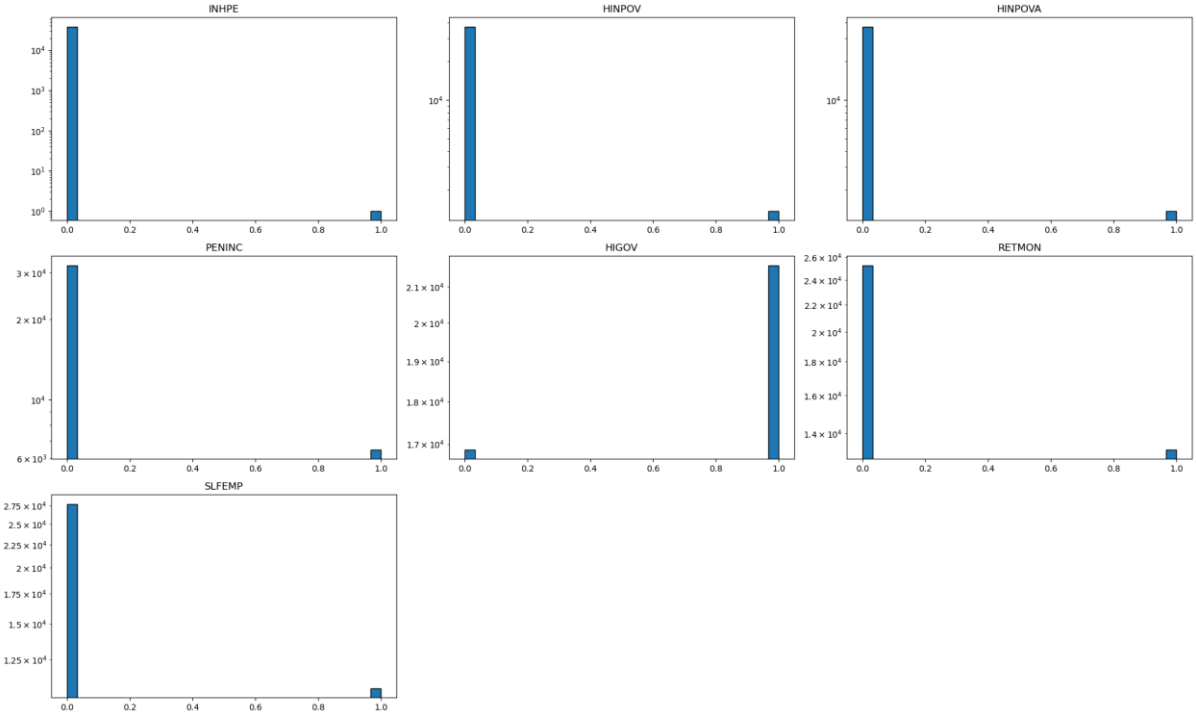

Figure 3

```
0,0,0,0,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.167408 -0.189673
MSTOT  -0.167408  1.000000  0.668407
COGTOT -0.189673  0.668407  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,0,0,0,1
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.168911 -0.189017
MSTOT  -0.168911  1.000000  0.685551
COGTOT -0.189017  0.685551  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,0,1,0,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.132131 -0.206216
MSTOT  -0.132131  1.000000  0.666644
COGTOT -0.206216  0.666644  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,0,1,0,1
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.098731 -0.188849
MSTOT  -0.098731  1.000000  0.586799
COGTOT -0.188849  0.586799  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0
```

Figure 4

```
0,0,1,1,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.114583 -0.178451
MSTOT  -0.114583  1.000000  0.620651
COGTOT -0.178451  0.620651  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,0,1,1,1
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.141674 -0.154367
MSTOT  -0.141674  1.000000  0.585759
COGTOT -0.154367  0.585759  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,1,1,0,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.107435 -0.150123
MSTOT  -0.107435  1.000000  0.608581
COGTOT -0.150123  0.608581  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

0,1,1,1,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.102097 -0.160233
MSTOT  -0.102097  1.000000  0.586923
COGTOT -0.160233  0.586923  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0


0,1,1,1,1
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.072136 -0.152807
MSTOT  -0.072136  1.000000  0.602385
COGTOT -0.152807  0.602385  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0

1,0,0,0,0
            SHLT     MSTOT    COGTOT
SHLT    1.000000 -0.085638 -0.155851
MSTOT  -0.085638  1.000000  0.727765
COGTOT -0.155851  0.727765  1.000000

Pearson correlation between SHLT and MSTOT: -0.14602879933827834, P-value: 3.1975319883446005e-159
Pearson correlation between SHLT and COGTOT: -0.18897019698202974, P-value: 3.618519293858185e-267
Pearson correlation between MSTOT and COGTOT: 0.6382783707627973, P-value: 0.0
```

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

Table 4

Group Pair	P-Value	Group Pair	P-Value
0,0,0,0,0 and 0,0,0,0,1	1.01E-12	0,0,1,0,0 and 0,1,1,1,1	1.95E-38
0,0,0,0,0 and 0,0,1,0,0	1.15E-36	0,0,1,0,0 and 1,0,0,0,0	1.27E-60
0,0,0,0,0 and 0,0,1,0,1	7.52E-73	0,0,1,0,1 and 0,0,1,1,0	9.75E-08
0,0,0,0,0 and 0,0,1,1,0	2.08E-81	0,0,1,0,1 and 0,0,1,1,1	7.05E-05
0,0,0,0,0 and 0,0,1,1,1	7.99E-71	0,0,1,0,1 and 0,1,1,0,0	0.103445
0,0,0,0,0 and 0,1,1,0,0	5.51E-20	0,0,1,0,1 and 0,1,1,1,0	0.108941
0,0,0,0,0 and 0,1,1,1,0	6.44E-61	0,0,1,0,1 and 0,1,1,1,1	0.005394
0,0,0,0,0 and 0,1,1,1,1	4.42E-52	0,0,1,0,1 and 1,0,0,0,0	2.12E-118
0,0,0,0,0 and 1,0,0,0,0	1.50E-109	0,0,1,1,0 and 0,0,1,1,1	0.014511
0,0,0,0,1 and 0,0,1,0,0	7.15E-39	0,0,1,1,0 and 0,1,1,0,0	6.76E-06
0,0,0,0,1 and 0,0,1,0,1	4.37E-44	0,0,1,1,0 and 0,1,1,1,0	6.66E-09
0,0,0,0,1 and 0,0,1,1,0	2.01E-57	0,0,1,1,0 and 0,1,1,1,1	2.80E-12
0,0,0,0,1 and 0,0,1,1,1	2.91E-53	0,0,1,1,0 and 1,0,0,0,0	3.35E-98
0,0,0,0,1 and 0,1,1,0,0	1.71E-19	0,0,1,1,1 and 0,1,1,0,0	0.000504
0,0,0,0,1 and 0,1,1,1,0	6.02E-42	0,0,1,1,1 and 0,1,1,1,0	0.000103
0,0,0,0,1 and 0,1,1,1,1	2.23E-42	0,0,1,1,1 and 0,1,1,1,1	1.33E-06
0,0,0,0,1 and 1,0,0,0,0	2.16E-101	0,0,1,1,1 and 1,0,0,0,0	9.56E-104
0,0,1,0,0 and 0,0,1,0,1	9.11E-41	0,1,1,0,0 and 0,1,1,1,0	0.647902
0,0,1,0,0 and 0,0,1,1,0	3.19E-23	0,1,1,0,0 and 0,1,1,1,1	0.069745
0,0,1,0,0 and 0,0,1,1,1	1.78E-29	0,1,1,0,0 and 1,0,0,0,0	3.94E-79
0,0,1,0,0 and 0,1,1,0,0	7.88E-17	0,1,1,1,0 and 0,1,1,1,1	0.021763
0,0,1,0,0 and 0,1,1,1,0	4.71E-39	0,1,1,1,0 and 1,0,0,0,0	7.94E-130
0,1,1,1,1 and 1,0,0,0,0	8.52E-117		

Table 6

	count	mean	std	min	25%	50%	75%	max
BMI	33538	1.41E-14	1.000015	-3.56846	-0.69774	-0.15828	0.535316	6.353814
INHPFN	33538	5.82E-15	1.000015	-0.12617	-0.12617	-0.12617	-0.12617	32.9837
HHHRES	33538	-4.92E-14	1.000015	-0.60029	-0.60029	-0.60029	0.366366	9.066287
HCHILD	33538	7.75E-15	1.000015	-1.72947	-0.65153	-0.11256	0.426414	4.199211
LIVSIB	33538	2.06E-15	1.000015	-1.21745	-0.79643	-0.3754	0.466642	6.360959
HAIRA	33538	-1.18E-14	1.000015	-0.41891	-0.41891	-0.41891	-0.05202	14.42363
HATOTB	33538	-1.00E-15	1.000015	-3.4569	-0.49971	-0.31643	0.089569	16.44927
IEARN	33538	-4.36E-14	1.000015	-0.77812	-0.77812	-0.36534	0.322617	10.22924
HITOT	33538	-1.34E-16	1.000015	-1.13878	-0.60941	-0.26853	0.285929	15.36991
PRPCNT	33538	8.69E-17	1.000015	-1.34082	-1.34082	0.358474	0.358474	37.74293
SHLT	33538	-4.80E-15	1.000015	-1.54866	-0.48538	-0.48538	0.577894	2.704447
MSTOT	33538	-2.33E-15	1.000015	-5.53478	-0.27574	0.308599	0.892937	0.892937
COGTOT	33538	-1.87E-16	1.000015	-4.66151	-0.54534	-0.03082	0.740961	2.799045

Table 8 Performance of Best Parameters

Group	Best Parameters	Train R^2 (avg)	Test R^2 (avg)	Train RMSE	Test RMSE	Train MAPE	Test MAPE
0,0,0,0,0	{'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}	0.692483	0.748598	0.016695	0.046585	0.003745	0.010571
0,0,0,0,1	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}	0.653973	0.790487	0.017321	0.042677	0.00384	0.010011
0,0,1,0,0	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}	0.682798	0.729988	0.017791	0.056201	0.003956	0.012163
0,0,1,0,1 + 0,1,1,0,0 + 0,1,1,1,0	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}	0.677	0.787799	0.016235	0.040685	0.003603	0.009468
0,0,1,1,0	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}	0.681261	0.731047	0.016531	0.046394	0.003701	0.010611
0,0,1,1,1	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}	0.674078	0.742469	0.017445	0.046525	0.003803	0.010676
0,1,1,1,1	{'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}	0.647207	0.816367	0.016529	0.037547	0.003643	0.008737
1,0,0,0,0	{'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}	0.524486	0.523801	0.025684	0.078733	0.006101	0.019442

Table 10

	0,0,1,1,1	*0,1,1,1,1*	*0,0,0,0,0*	*0,0,0,0,1*	*0,0,1,0,0*	*0,0,1,0,1 + 0,1,1,0,0 + 0,1,1,1,0*	*0,0,1,1,0*	*1,0,0,0,0*
BMI	0.206846	0.205366	0.186807	0.206882	0.177118	0.182868	0.165806	0.190242
HAIRA	0.102273	0.165408	0.063827	0.103278	0.080345	0.100472	0.093817	0.00816
HATOTB	0.211099	0.164281	0.18522	0.204477	0.194356	0.204485	0.215859	0.187186
HCHILD	0.089465	0.086882	0.068439	0.07382	0.078711	0.07752	0.081915	0.123934
HHHRES	0.027824	0.018887	0.048396	0.045222	0.0331	0.025929	0.020001	0.07556
HITOT	0.197395	0.190729	0.197449	0.179372	0.186548	0.192284	0.181478	0.154372
IEARN	0.041733	0.041454	0.139575	0.06601	0.132384	0.102273	0.122888	0.08602
INHPFN	0.009942	0.011675	0.004504	0.001703	0.004482	0.002776	0.005638	0.001495
LIVSIB	0.085224	0.090251	0.080669	0.094715	0.081249	0.084104	0.085034	0.146708
PRPCNT	0.0282	0.025067	0.025113	0.024522	0.031707	0.027289	0.027564	0.026324