**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 |
| **Department** | **:** | McKelvey School of 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  0 | 0  1  1 | 1  1  1 | 0  0  1 | 1  0  0 | 6138 |
| 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.

To mitigate the potential adverse effects of disparate ranges across continuous features on machine learning models and performance metrics, standardization was employed for these features. This process ensures that each feature contributes equally to the model's prediction capability by normalizing their ranges. The statistical characteristics of the standardized features are comprehensively detailed in Table 6 of the Appendix, illustrating the distribution and scale uniformity post-standardization.

### 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 small 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, two alternative data preparation strategies were devised. The first strategy (abbreviate as dataset 1) involves excluding all data samples containing negative values for "HATOTB" and subsequently applying the logarithmic transformation across all remaining features. The second strategy (abbreviate as dataset 2) retains all samples but restricts the logarithmic transformation to features excluding "HATOTB".

The efficacy of the model will be evaluated across these two distinct dataset versions to determine the impact of each data preparation strategy on model performance. This comparative analysis will facilitate a deeper understanding of the influence of data preprocessing on the predictive accuracy and reliability of the model.

The descriptive statistics of dataset with first strategy is shown in Table 7 below.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| SHLT | 32145 | 0.855707 | 0.407729 | 0.09531 | 0.741937 | 0.741937 | 1.131402 | 1.629241 |
| BMI | 32145 | 3.325061 | 0.175121 | 2.282382 | 3.202746 | 3.314186 | 3.433987 | 4.11578 |
| MSTOT | 32145 | 2.600543 | 0.139179 | 1.410987 | 2.572612 | 2.646175 | 2.714695 | 2.714695 |
| COGTOT | 32145 | 3.174848 | 0.170786 | 1.808289 | 3.095578 | 3.182212 | 3.299534 | 3.558201 |
| INHPFN | 32145 | -2.26132 | 0.315842 | -2.30259 | -2.30259 | -2.30259 | -2.30259 | 1.629241 |
| HHHRES | 32145 | 0.940872 | 0.304001 | 0.741937 | 0.741937 | 0.741937 | 1.131402 | 2.493205 |
| HCHILD | 32145 | 0.98035 | 0.8099 | -2.30259 | 0.741937 | 1.131402 | 1.410987 | 2.406945 |
| LIVSIB | 32145 | 0.628872 | 1.19165 | -2.30259 | 0.09531 | 0.741937 | 1.410987 | 2.895912 |
| HAIRA | 32145 | 4.143813 | 6.685256 | -2.30259 | -2.30259 | -2.30259 | 11.0021 | 14.61527 |
| HATOTB | 32145 | 12.13331 | 2.00646 | -2.30259 | 11.42954 | 12.38003 | 13.24281 | 16.41039 |
| IEARN | 32145 | 6.810352 | 5.491602 | -2.30259 | -2.30259 | 9.615812 | 10.59664 | 12.89922 |
| HITOT | 32145 | 11.12746 | 1.052774 | -2.30259 | 10.67126 | 11.17049 | 11.66135 | 14.10069 |
| PRPCNT | 32145 | -0.51623 | 1.099296 | -2.30259 | -2.30259 | 0.09531 | 0.09531 | 3.139833 |

Table 7

# 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 8.

|  |  |  |
| --- | --- | --- |
| 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 8

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's 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

Random Forest Regressor is trained separately for each group on both dataset 1 and 2. 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 9 and 10 in the Appendix present the optimal parameters for each group across different datasets, along with their corresponding performance metrics. These tables indicate that, overall, the Random Forest Regressor achieves superior performance with dataset 2. However, the accuracy that is measured by RMSE in predicting the groups labeled “0,0,1,1,1” and “0,1,1,1,1” is higher when using dataset 1. Consequently, the decision has been made to train the models for these specific groups on dataset 1, while employing dataset 2 for the remaining groups. This approach is intended to leverage the strengths of each dataset to enhance model performance across different group classifications.

## 3.4 Prescriptive methods

## 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 decide whether a point is outlier.

**ML2**

ML2 is the process of Standardization. By subtracting the mean and dividing by the standard deviation for each value of each feature, the range of all features are transformed to same scale.

**ML3**

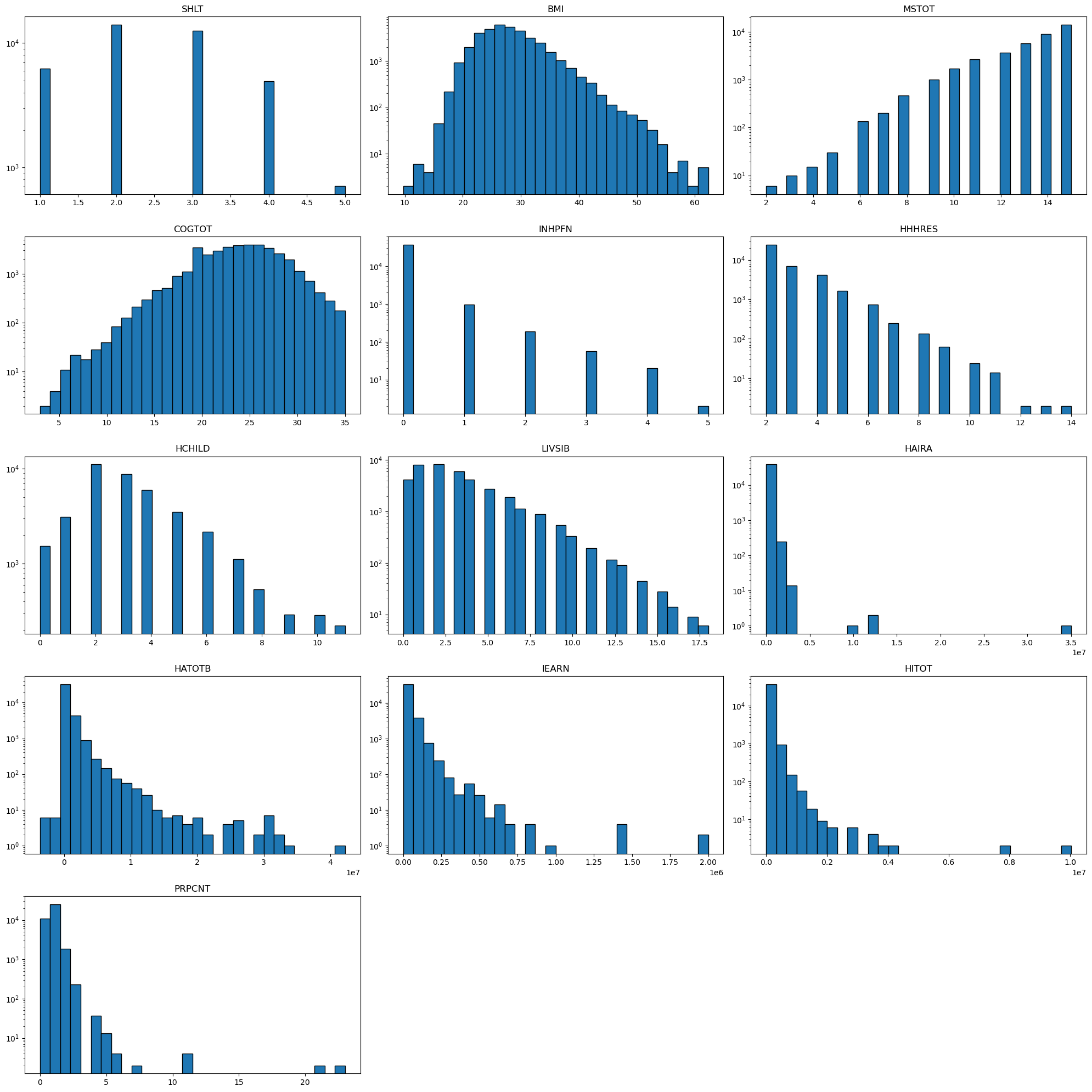
ML3 is the Random Forest Regressor. It takes predictive features and predict response features by the average of every tree’s output. It does not require a prior distribution but do need MSE as loss function to optimize the result.

# Part 6. Source Code

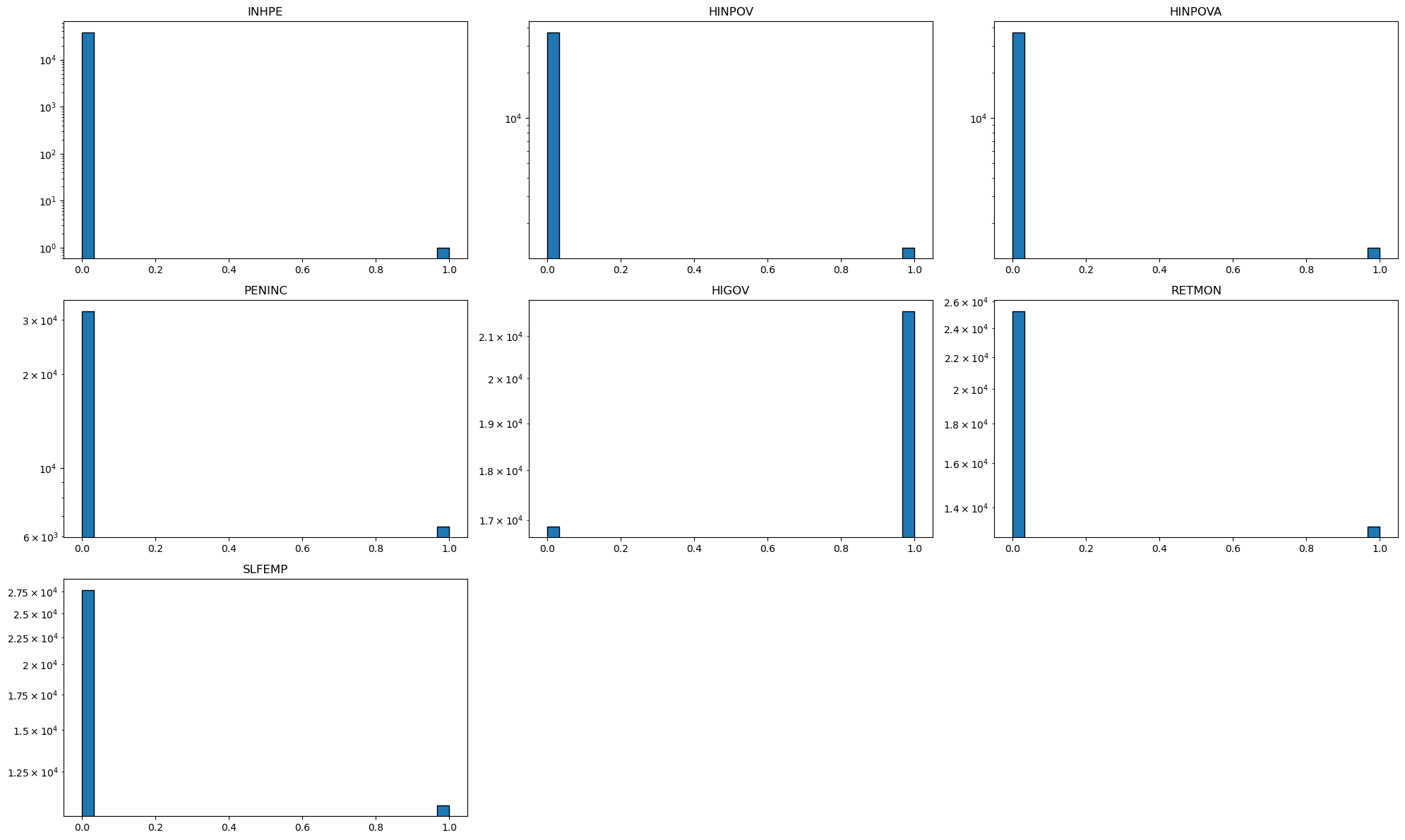
<https://github.com/YuzhenZhou1327/ESE527_Project_HRS>

# Part 7. Appendix

### Figure 1 Histogram of Numerical Ordinal Features



### Figure 2 Histogram of Binary Categorical Features



### Figure 3

文本

描述已自动生成

### Figure 4

文本

低可信度描述已自动生成文本

描述已自动生成

### Table

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **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 9 Performance of Best Parameters on [Dataset 1](#_Feature_Transformation)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.691721 | 0.752967 | 0.050075 | 0.142923 | 0.085123 | 0.257604 |
| 0,0,0,0,1 | {'max\_depth': 30, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.652693 | 0.791388 | 0.055133 | 0.135515 | 0.106353 | 0.287286 |
| 0,0,1,0,0 | {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.680689 | 0.725043 | 0.048591 | 0.140366 | 0.071974 | 0.214615 |
| 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} | 0.679475 | 0.79 | 0.047922 | 0.127269 | 0.076108 | 0.212265 |
| 0,0,1,1,0 | {'max\_depth': 30, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.67986 | 0.735613 | 0.047316 | 0.134255 | 0.068566 | 0.190829 |
| 0,0,1,1,1 | {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.672306 | 0.754941 | 0.04983 | 0.133109 | 0.073377 | 0.187129 |
| 0,1,1,1,1 | {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.662784 | 0.825263 | 0.049739 | 0.106976 | 0.074897 | 0.137458 |
| 1,0,0,0,0 | {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.503717 | 0.483642 | 0.068512 | 0.242483 | 0.091798 | 0.381998 |

### Table 10 Performance of Best Parameters on [Dataset 2](#_Feature_Transformation)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.685567 | 0.776914 | 0.049753 | 0.129077 | 0.08562 | 0.2093 |
| 0,0,0,0,1 | {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.646708 | 0.767416 | 0.054531 | 0.124954 | 0.108847 | 0.232312 |
| 0,0,1,0,0 | {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.701678 | 0.769488 | 0.048522 | 0.129458 | 0.071274 | 0.201153 |
| 0,0,1,0,1 + 0,1,1,0,0 + 0,1,1,1,0 | {'max\_depth': 30, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.68564 | 0.781457 | 0.04807 | 0.121976 | 0.077237 | 0.212279 |
| 0,0,1,1,0 | {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.677022 | 0.762959 | 0.04718 | 0.121304 | 0.068142 | 0.164933 |
| 0,0,1,1,1 | {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.688858 | 0.734571 | 0.047765 | 0.139326 | 0.070469 | 0.201918 |
| 0,1,1,1,1 | {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.68044 | 0.75027 | 0.048903 | 0.117502 | 0.072609 | 0.193155 |
| 1,0,0,0,0 | {'max\_depth': 20, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 300} | 0.466553 | 0.588631 | 0.073742 | 0.185971 | 0.101027 | 0.212058 |