Supplement to Example Manuscript Template for a Data Analysis Project

Sri Lakshmi Chundru, Collin Real, Joaquin Ramirez, Seth Harris

1. Overview

An overview of the correct process to generate the final outputs as well as some exploratory analysis.

2. Code Information

Run analysis in the following order.

1. Pre-Processing File (Quarto): /code/processing-code/pre_process_tax_data.qmd - preprocesses the tax/ev data and merges them into one dataset.
2. Exploratory Data Analysis (Quarto): /code/eda-code/eda-ev-dataset.qmd - exploratory analysis of the electric vehicle dataset /code/eda-code/eda-ev-tax.qmd - exploratory analysis of the tax dataset
3. Predictive models (Quarto): /code/analysis-code/statistical_analysis.qmd - builds multiple predictive models, visualizes, and compares them.

3. Additional Method Details

3.0.0.1 Decision Tree Regressor:

A basic model that splits the data into subsets based on feature values, creating a tree structure. We set hyperparameters such as the maximum depth and minimum samples per leaf to control the model's complexity and prevent overfitting.

3.0.0.2 Random Forest Regressor:

An ensemble learning method that constructs multiple decision trees and aggregates their outputs to improve prediction accuracy. We configured key parameters like the number of trees (n_estimators) and the maximum depth, optimizing for a balance between bias and variance.

3.0.0.3 XGBoost Regressor:

A gradient boosting model known for its efficiency and performance. XGBoost builds trees sequentially, with each new tree aiming to correct errors made by the previous ones. We

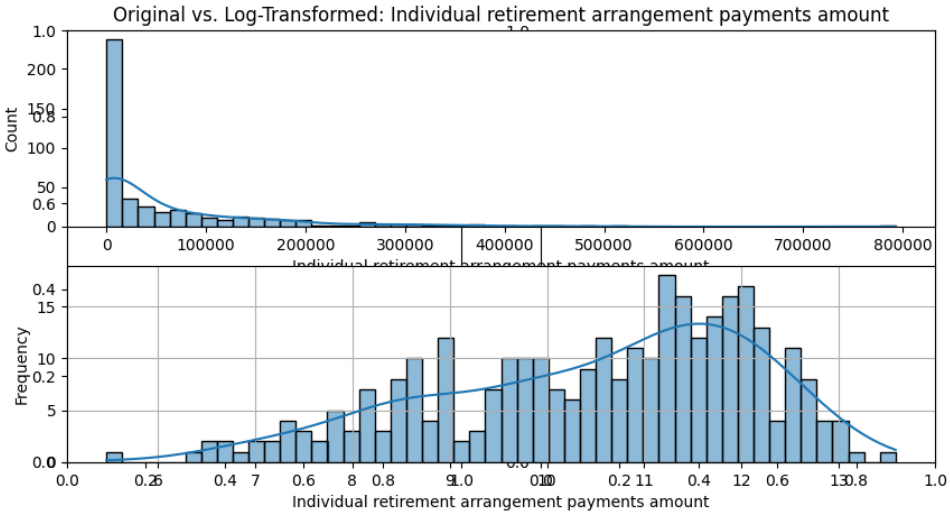
fine-tuned parameters such as the learning rate (eta), maximum depth, and regularization terms (lambda and alpha) to enhance model performance.

3.0.0.4 HistGradientBoosting Regressor:

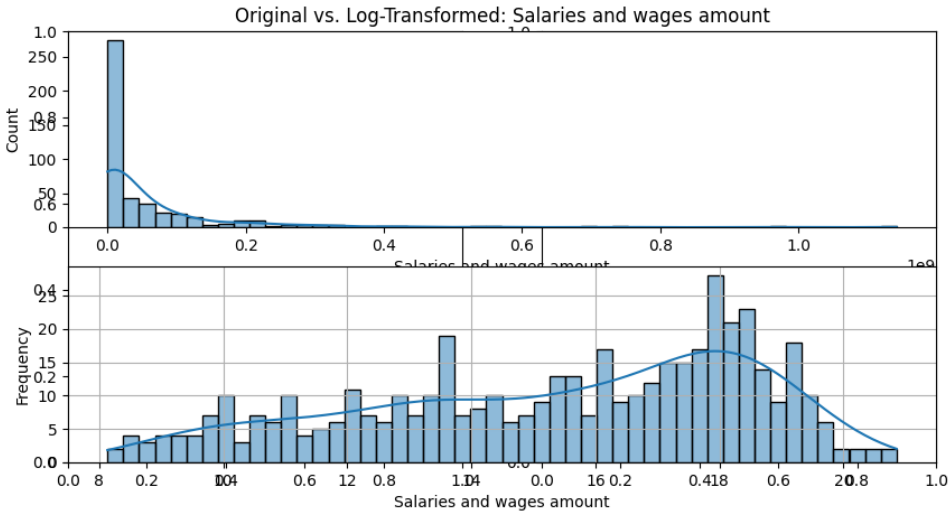
A histogram-based implementation of gradient boosting that is particularly efficient for large datasets. This model discretizes continuous features into bins, reducing memory usage and speeding up training. We set parameters including the number of iterations (max_iter), learning rate, and number of bins (max_bins) to optimize the model.

4. Additional results

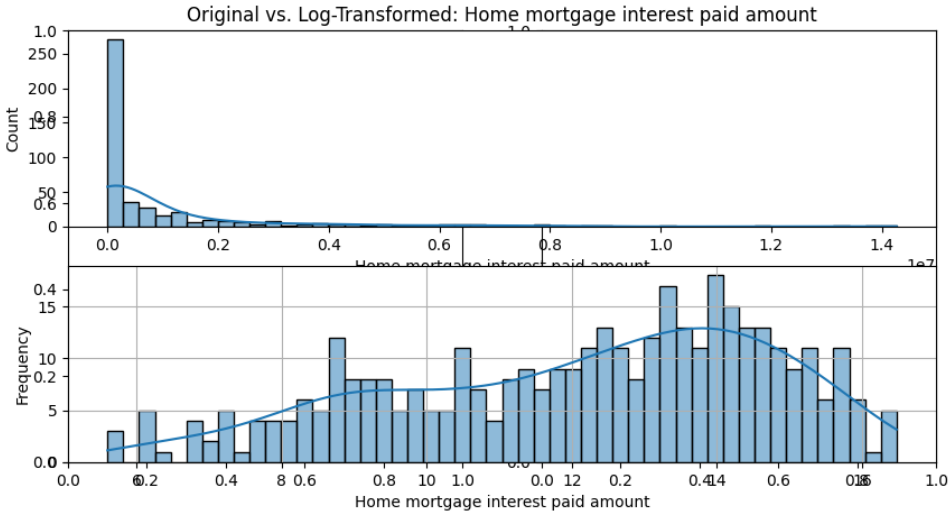
4.1 Log-Transformed Histogram Examples



Individual Retirement Payments

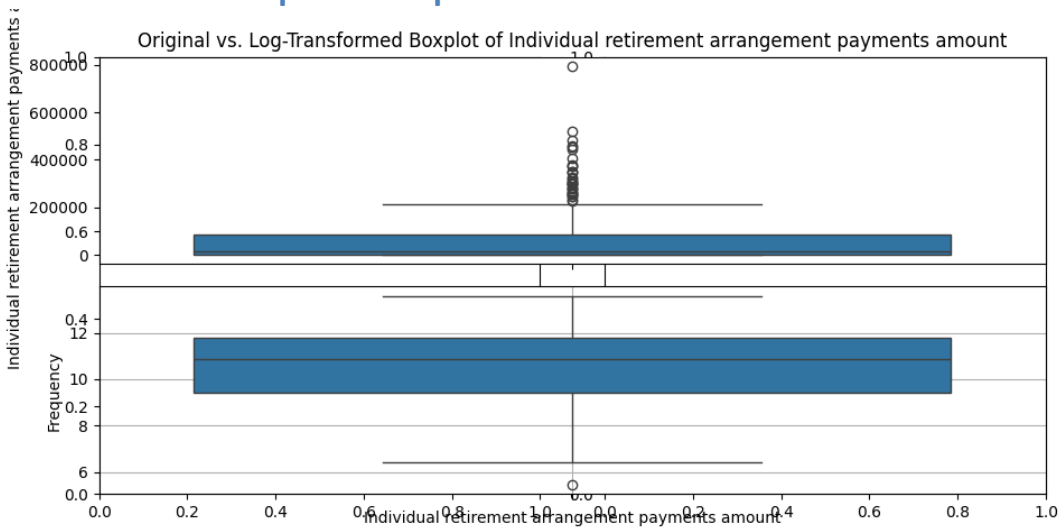


Salary & Wages Amount



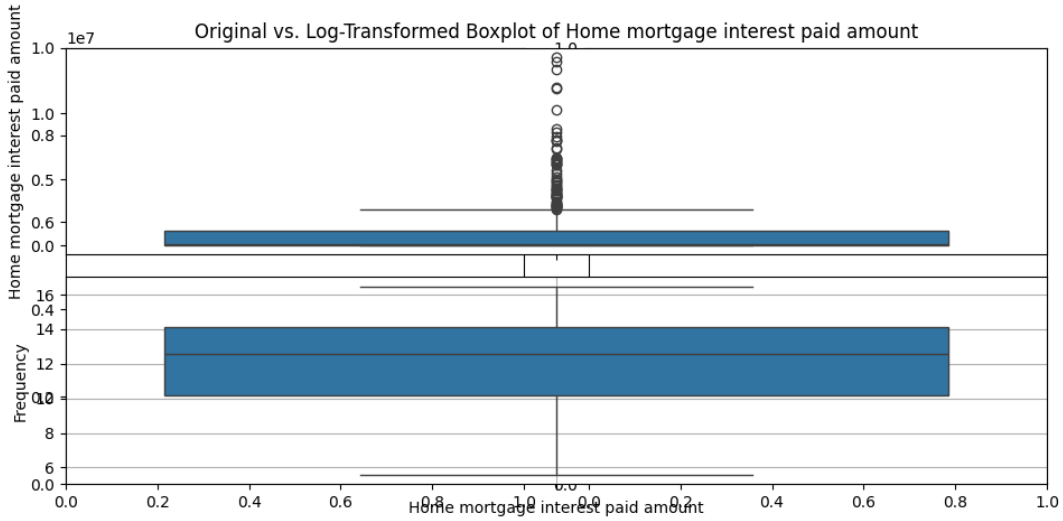
Home Mortgage Interest Paid

4.2 Log-Transformed Boxplot Examples



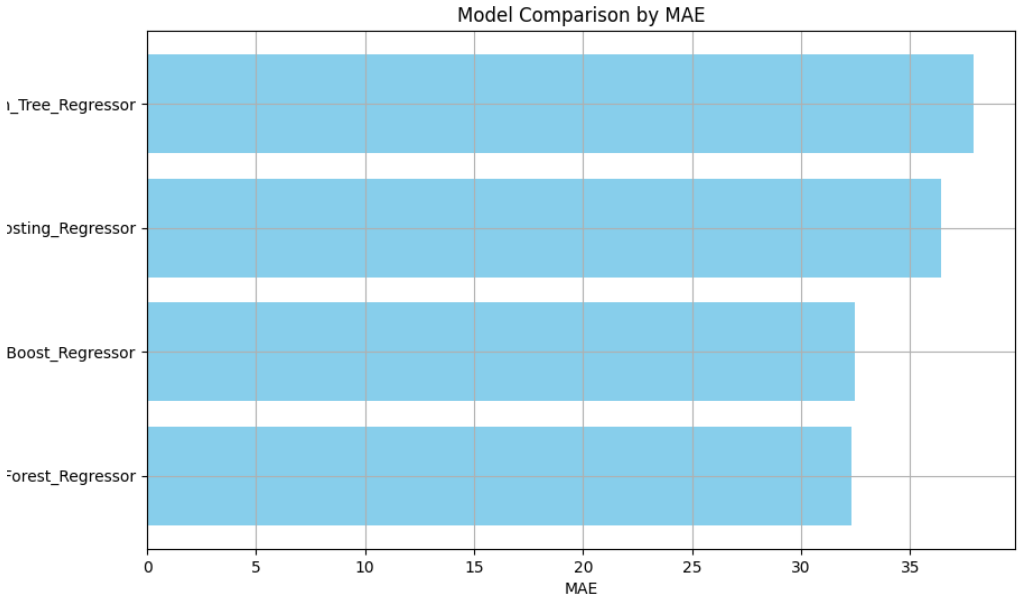
Individual Retirement Payments

Salary & Wages Amount



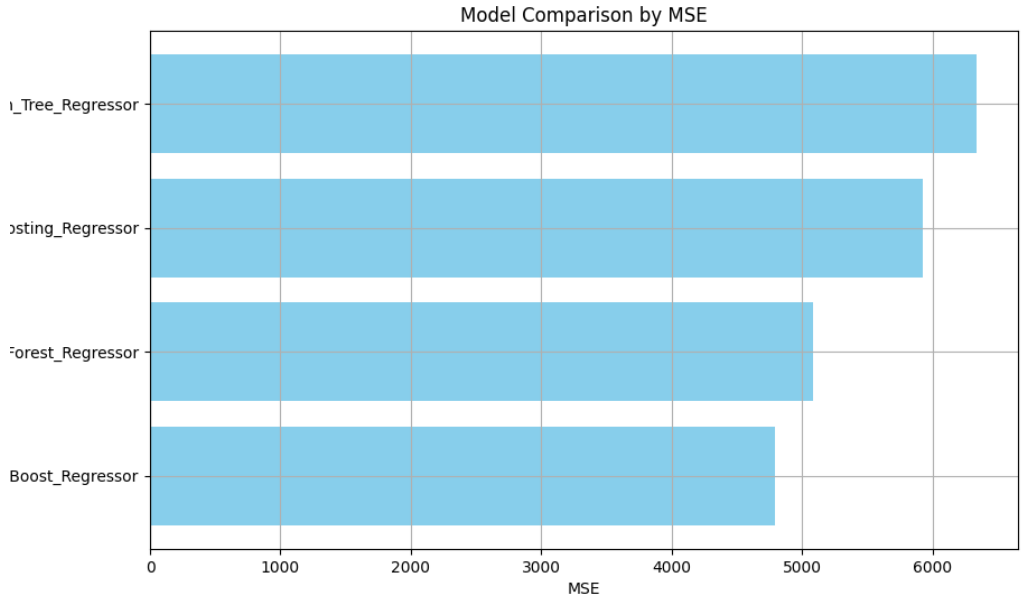
Home Mortgage Interest Paid

4.3 Model Comparison - MAE



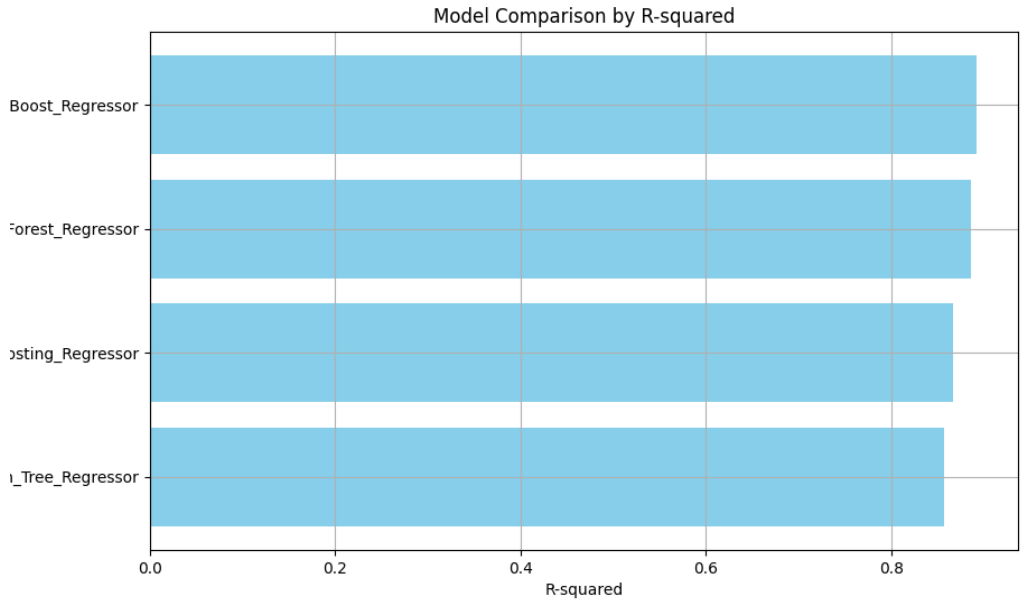
'MAE'

4.4 Model Comparison - MSE



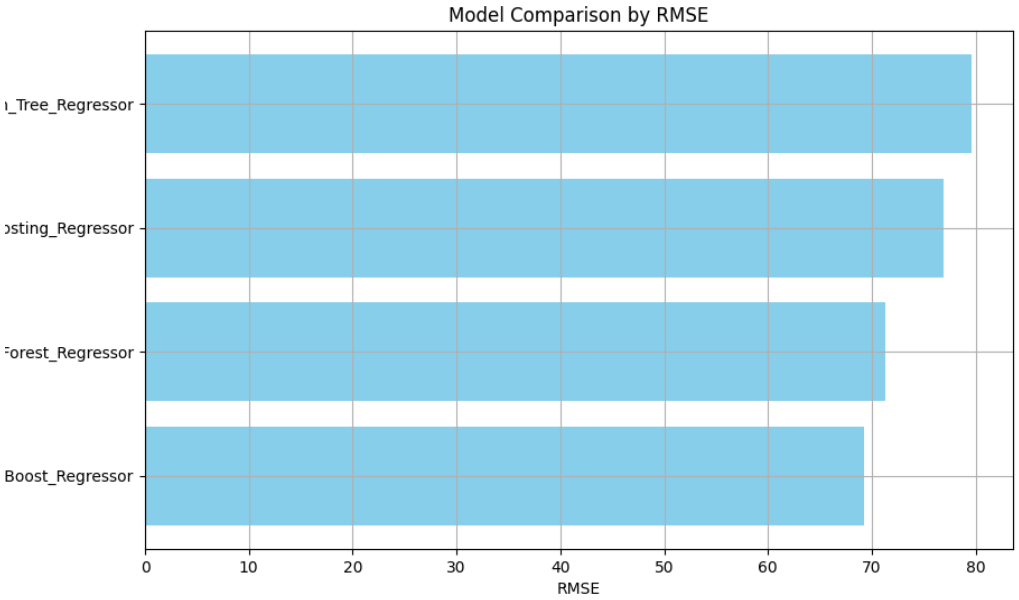
'MSE'

4.5 Model Comparison - R-squared



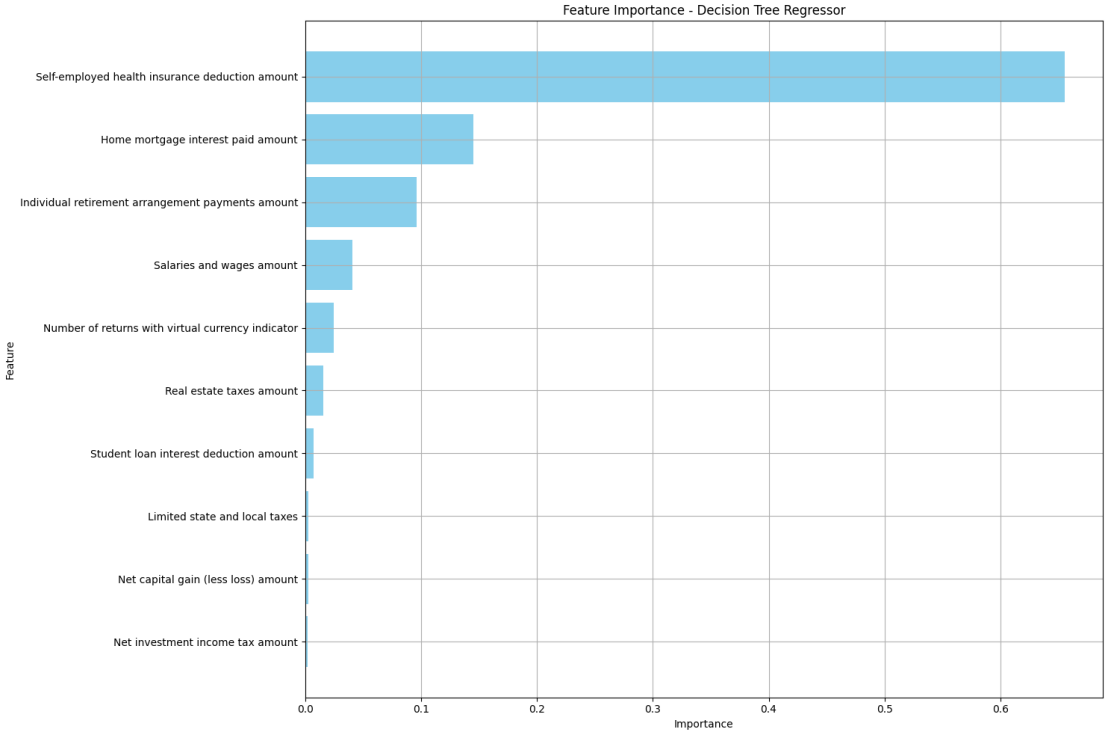
'R-squared'

4.6 Model Comparison - RMSE



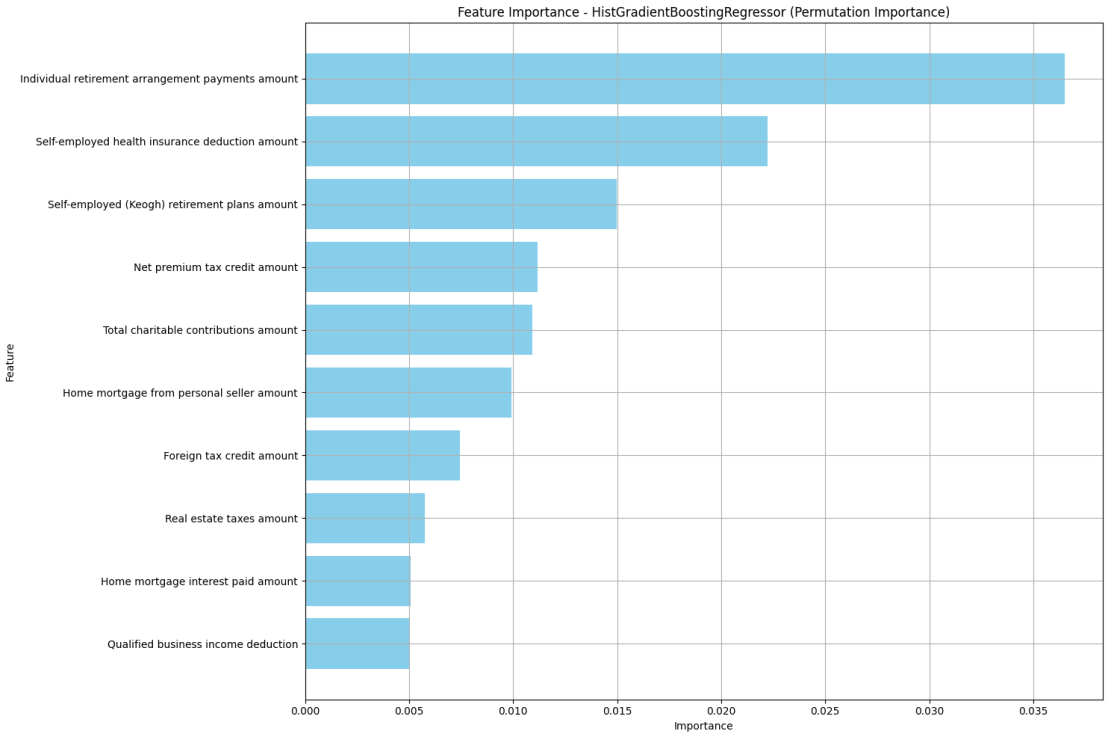
'RMSE'

4.7 Variable Importance - Decision Tree



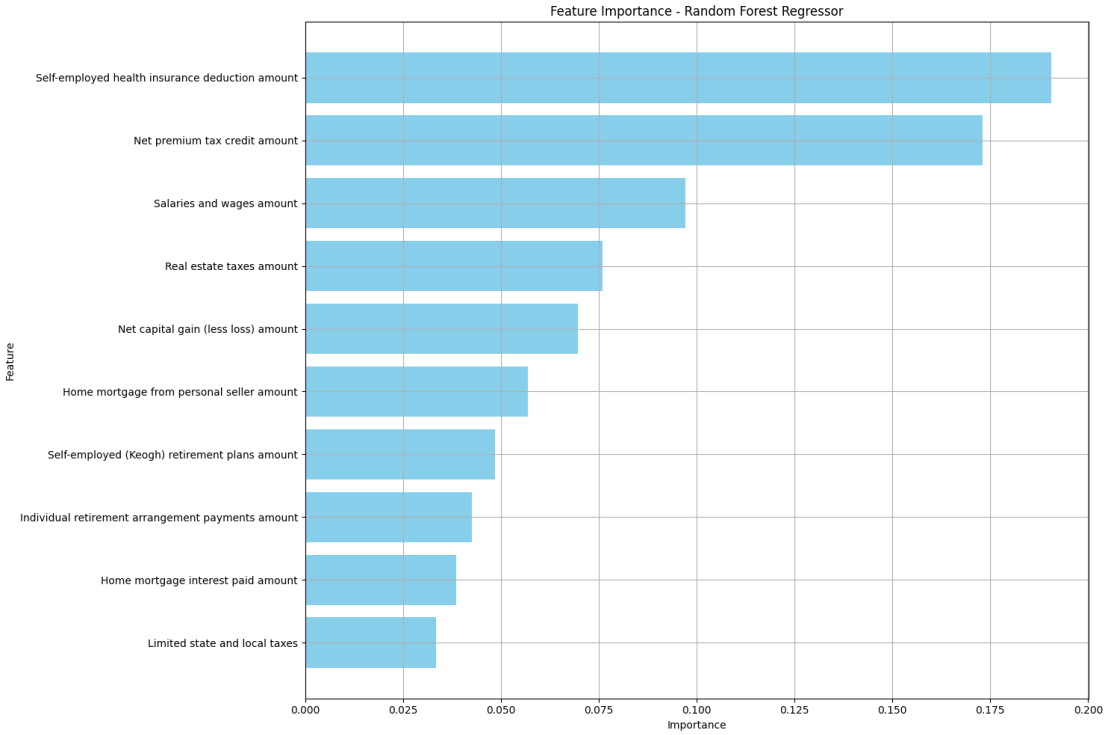
'Decision Tree Regressor'

4.8 Variable Importance - Hist Gradient Boosting



'Hist Gradient Boosting'

4.9 Variable Importance - Random Forest



'Random Forest'

5. References

- Jain, A. (n.d.). Electric vehicle population [Data set]. Kaggle. Retrieved June 6, 2024, from https://www.kaggle.com/datasets/jainaru/electric-vehicle-population
- Internal Revenue Service. (n.d.). SOI tax stats - Individual income tax statistics - ZIP code data (SOI). Retrieved June 16, 2024, from https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi