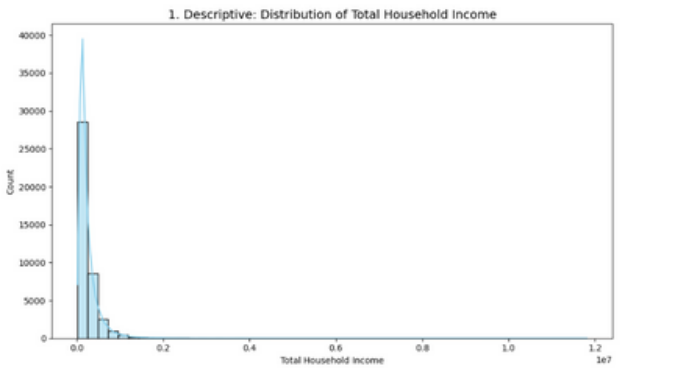
FAMILY INCOME AND EXPENDITURE

DISTRIBUTION OF THE TARGET VARIABLE (INCOME)

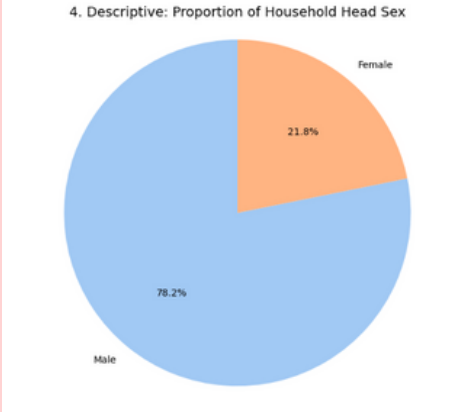


Target income is highly right-skewed; most households have low income, suggesting a non-linear model is appropriate.

PROPORTION OF HOUSEHOLD HEAD SEX (PIE CHART)

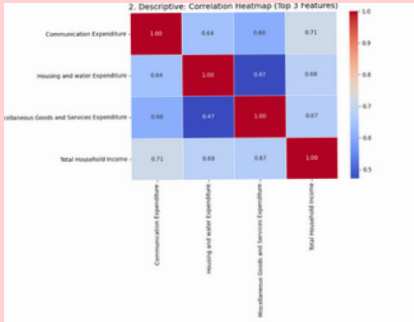
FEMALE

MALE



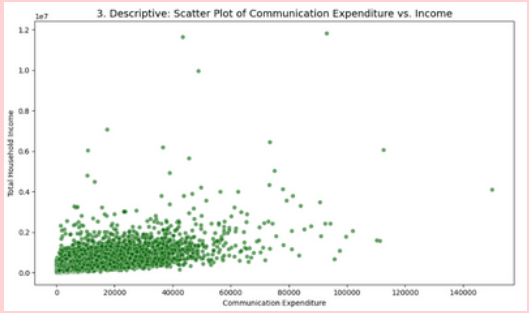
The pie chart clearly illustrates the proportional distribution of household head gender, indicating the relative frequency of male versus female household heads in the dataset. Most households are headed by males.

CORRELATION HEATMAP OF TOP 3 FEATURES



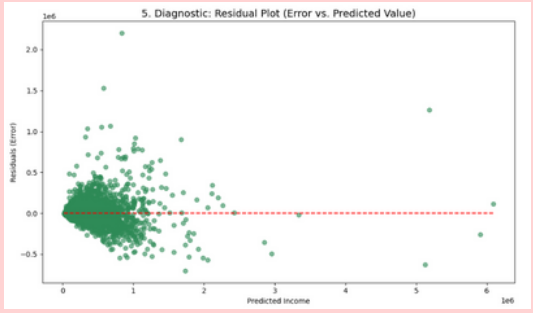
Examining the heatmap shows strong positive correlation between key expenditure categories (like 'Total Food Expenditure') and 'Total Household Income'. This confirms consumption is highly dependent on income levels

SCATTER PLOT OF TOP CORRELATED FEATURE VS. INCOME



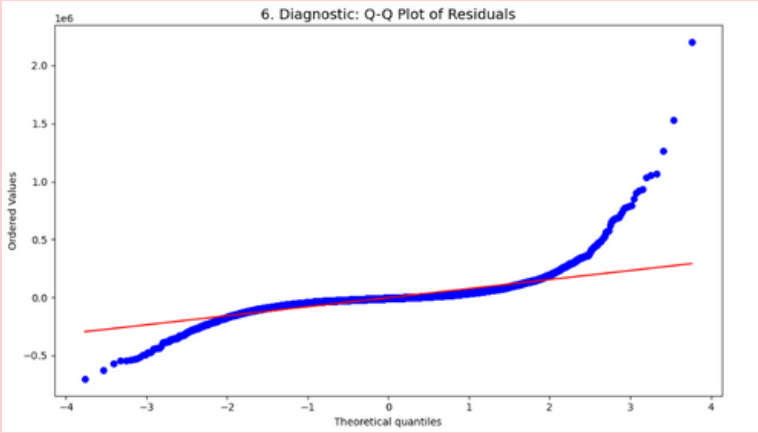
The scatter plot reveals a clear linear relationship between {top_feature} and income, indicating that households with higher food expenditure generally have higher incomes. This feature will be highly influential in the model.

RESIDUAL PLOT (HOMOSCEDASTICITY CHECK)



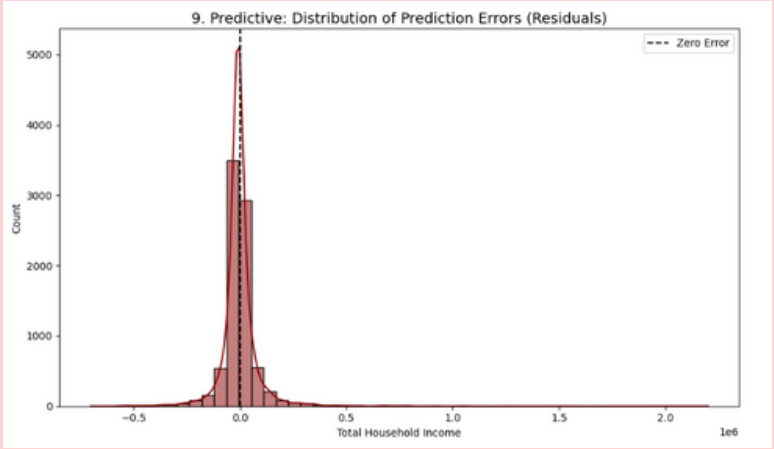
The residual plot for Random Forest generally shows a reduced fanning pattern compared to Linear Regression, indicating better handling of the variance across different income levels, though error variance remains highest for the highest predicted incomes.

Q-Q PLOT OF RESIDUALS (NORMALITY CHECK



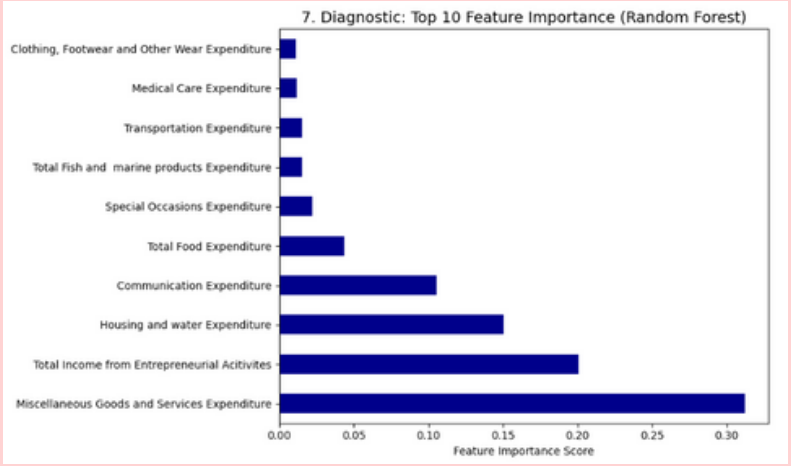
The Q-Q plot still shows non-normal residuals, especially in the tails. While Random Forest doesn't assume residual normality, this plot highlights that the largest prediction errors deviate significantly from a normal distribution.

ERROR DISTRIBUTION



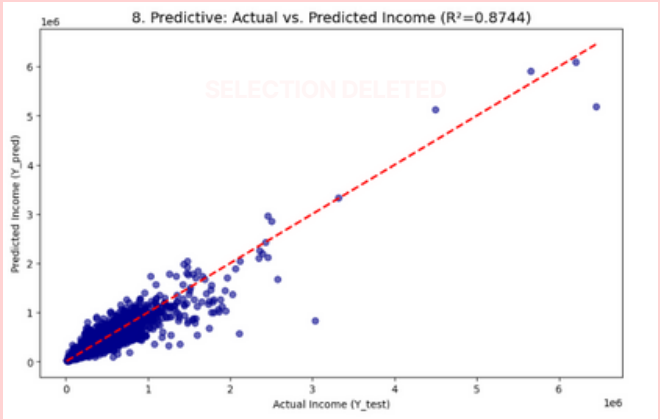
The error distribution is narrower and more centered around zero compared to the Linear Regression model, indicating smaller overall prediction errors. The improved symmetry suggests the Random Forest model is more robust against the high-income outliers

FEATURE IMPORTANCE (MODEL IMPORTANCE SCORES)



Feature Importance scores from Random Forest clearly identify the most critical variables (likely expenditure categories) driving the income prediction, confirming their non-linear contribution to the model's performance.

ACTUAL VS. PREDICTED INCOME (PREDICTION QUALITY)



Compared to Linear Regression, the Random Forest model shows a tighter clustering of points around the ideal 45-degree line, particularly at higher income values, reflecting its superior ability to handle non-linear relationships in the data.