Urban vs Rural: A Machine Learning Approach to Household Classification in the Philippines

by Rachel Joy Baldo

Significance of the Study

Using machine learning to classify Philippine households as Urban or Rural based on their socioeconomic factors can help reveal fundamental differences. By doing so, policymakers, researchers, and social scientists can design more holistic programs, laws, and interventions to benefit both grassroots of urban and rural communities.

Objective

This study explores how income, expenditure, and consumption habits determine the classification of Philippine households as Urban or Rural: using the 2021 Family Income and Expenditure Survey (FIES) dataset provided by the Philippine Statistics Authority. Utilizing machine learning methods, the study hopes to uncover the factors that influence these economic behaviors and the urban-rural divide.

Problem Statement

Is it possible to predict households as Urban or Rural based on six variables: family size, per capita income, total expenditure, food expenditure, income, and non-food expenditure?

How well can six FIES variables: Family Size, Per Capita Income, Total Expenditure, Total Food Expenditure, Total Income, and Total Non-Food Expenditure, classify households as Urban(0) or Rural(1)?

Scope and Limitations

The research focuses on six FIES variables as features: Family Size, Per Capita Income, Total Expenditure, Food Expenditure, Income, and Non-Food Expenditure and are defined as follows:

- Family Size (FSIZE): the number of people in a household, which greatly influences its consumption patterns and financial priorities.
- Per Capita Income (PCINC): the total income divided by the family size of a household which is a measure of its economic status.

Scope and Limitations

- Total Expenditure (TTOTEX): the amount a household spends on food and non-food items, which provides a comprehensive view of its spending.
- Total Food Expenditure (TFOOD): the amount a household spends on food, reflecting its cost of living, dietary preferences, and food security and nutrition levels.
- Total Income (TOINC): the amount a household makes from all income sources, such as wages, business profits, and remittances, which measures its financial resources precisely.

Scope and Limitations

- Total Non-Food Expenditure (TNFOOD): the amount a household pays for other expenses apart from food, e.g., housing, utilities, education, and healthcare, offering insight into its resource allocation to different needs.
- Urban or Rural (URB): the target variable that classifies households by their urban(0) or rural(1) location, influencing their lifestyle and economic opportunities.

Methodology

Data Gathering

PSA's Family Income and 2021 Expenditure Survey Volume 2 data

- Family Size (FSIZE)
- Per Capita Income (PCINC)
- Total Expenditure (TTOTEX)
- Total Food Expenditure (TFOOD)
- Total Income (TOINC)
- Total Non-Food Expenditure (TNFOOD)
- Urban or Rural (URB)

Data Preprocessing

- Exploratory Analysis
- Descriptive Analysis
- Outlier Visualization
- Checking Class Imbalance
- Changing binary legends
- Normalization of data

Machine Learning

- Data splitting - 80% training set and 20% test set
- 10-fold Cross Validation
- Grid Search
- Model Testing
- Hyperparameter Tuning
- Evaluation Metrics
  - Precision
  - Recall
  - F1-Score
  - Accuracy

Post-hoc Analysis

- Confusion Matrix
- ROC-AUC Curve
- Feature Importance Analysis
  - SHAP values
    - Bar graph
    - Density Scatter plot
    - Force plot

Results and Findings

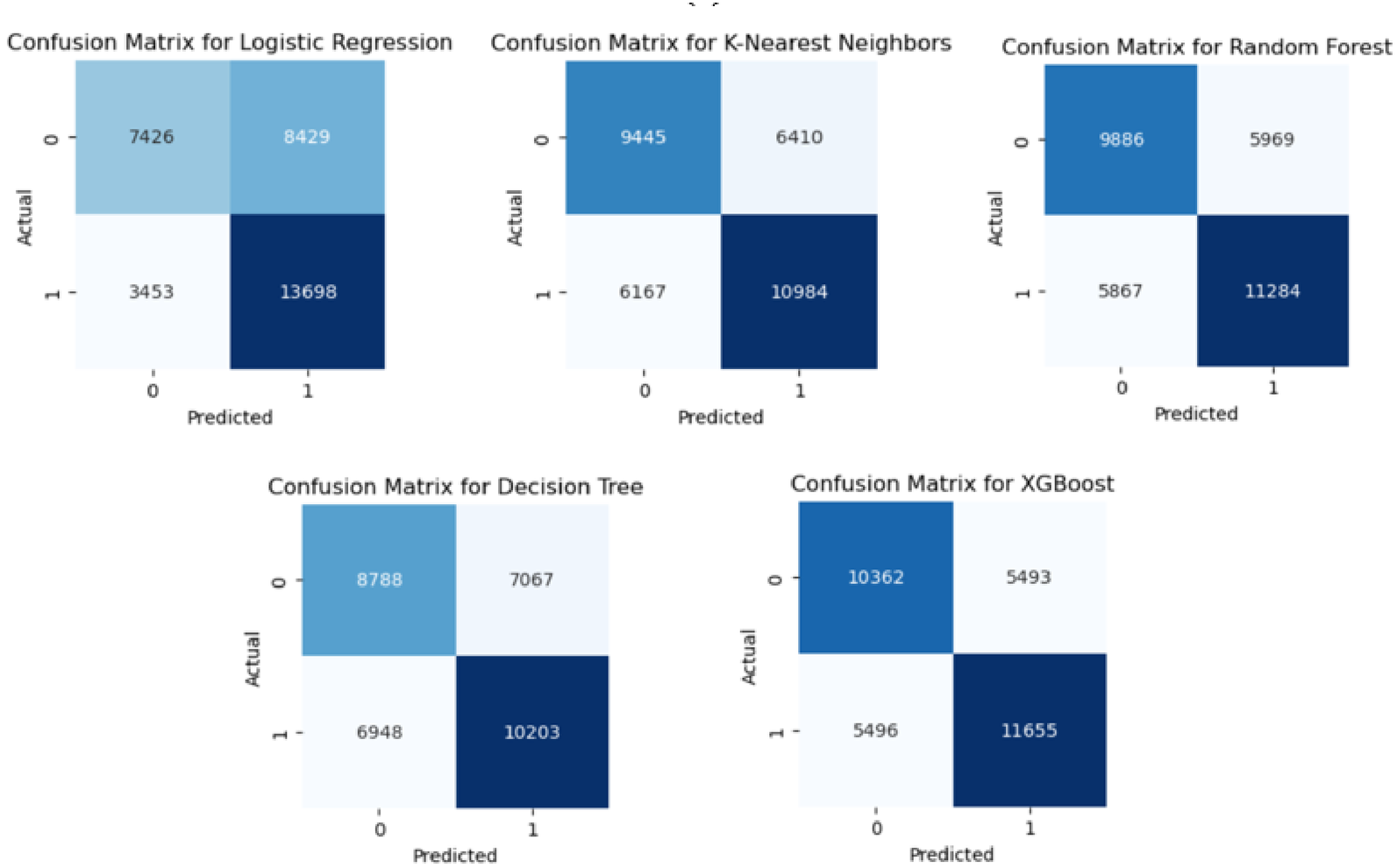
3.1 Machine Learning Modeling Results

Table 1. Summary of the best validation performance per ML model after 10-fold cross validation and grid search.

Class	ML Model	Validation Performance				Best Parameters
		Precision	Recall	F1-Score	Accuracy	
0 - Urban	Logistic Regression	0.68	0.47	0.56	0.64	C': 100, 'penalty': 'l2', 'solver': 'liblinear'
	KNN	0.63	0.62	0.62	0.64	'n_neighbors': 11
	RF	0.65	0.66	0.66	**0.67**	n_estimators': 100, 'min_samples_split': 5, 'min samples leaf': 2, 'max depth':
	Decision Tree	0.64	0.66	0.65	0.66	max_depth': 10, 'min_samples_leaf': 1, 'min samples split': 5
	XGBoost	0.66	0.66	0.66	**0.67**	learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 500
1 - Rural	Logistic Regression	0.62	0.80	0.70	0.64	C': 100, 'penalty': 'l2', 'solver': 'liblinear'
	KNN	0.65	0.66	0.66	0.64	'n_neighbors': 11
	RF	0.68	0.67	0.68	**0.67**	n_estimators': 100, 'min_samples_split': 5, 'min samples leaf': 2, 'max depth':
	Decision Tree	0.68	0.66	0.67	0.66	max_depth': 10, 'min_samples_leaf': 1, 'min samples split': 5
	XGBoost	0.68	0.68	0.68	**0.67**	learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 500

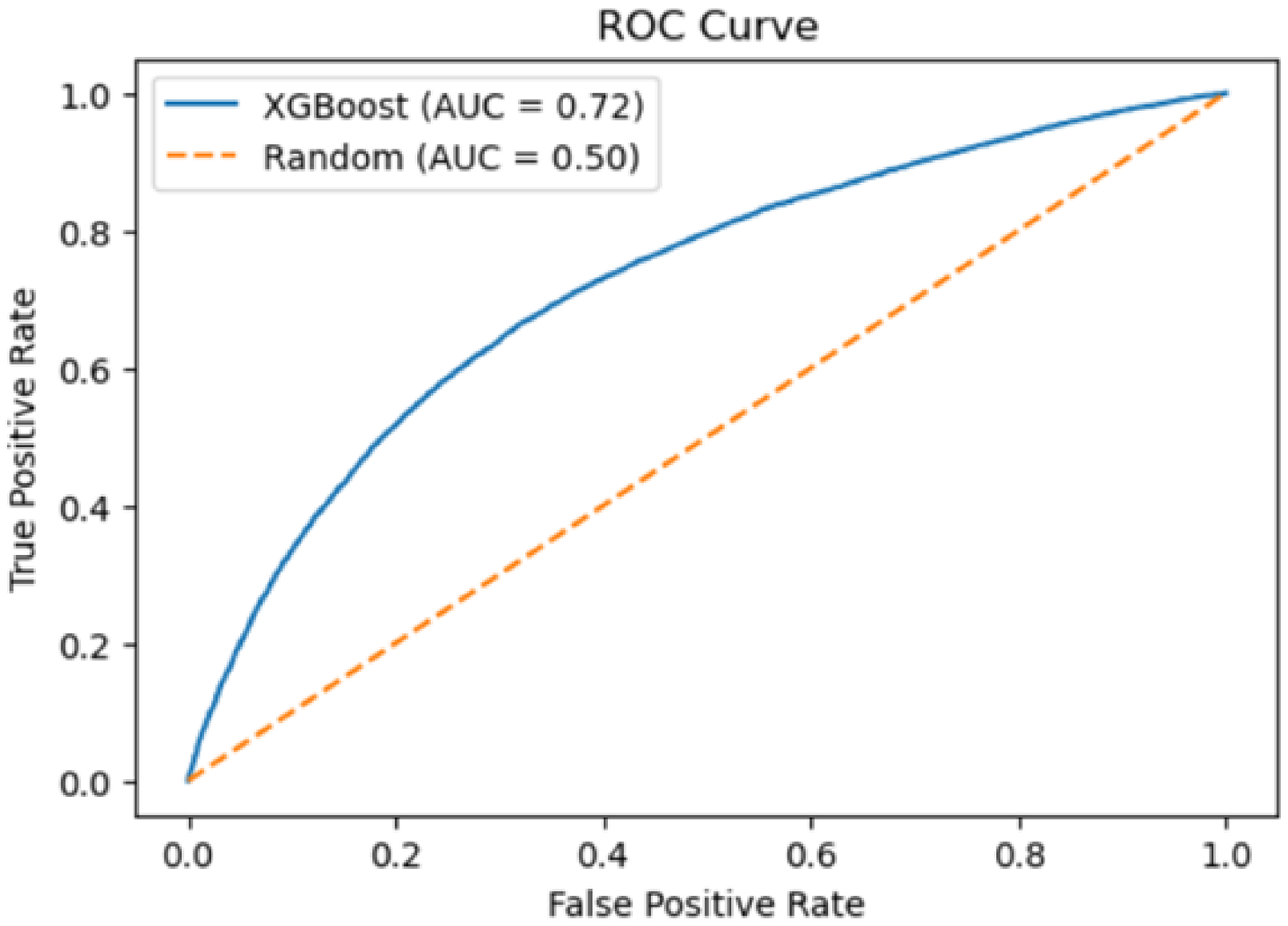
Results and Findings

Figure 1b. Confusion Matrix for each classification model



Results and Findings

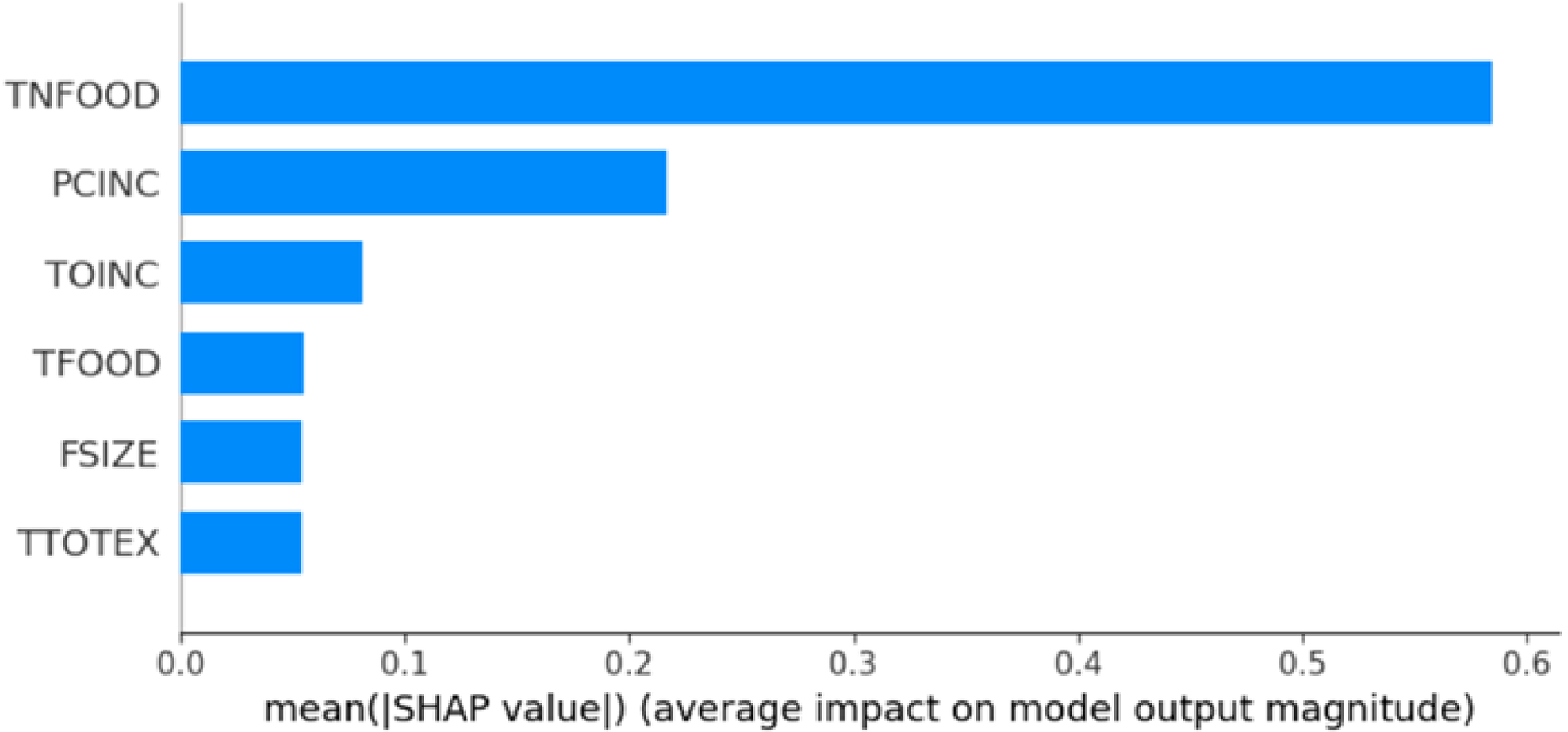
Figure 2. Area Under the Receiver Operating Characteristic (ROC-AUC) Curve for XGBoost model.



Results and Findings

3.2 Feature Importance Analysis

Figure 3. Bar graph of mean SHAP values for the features in XGBoost classifier.

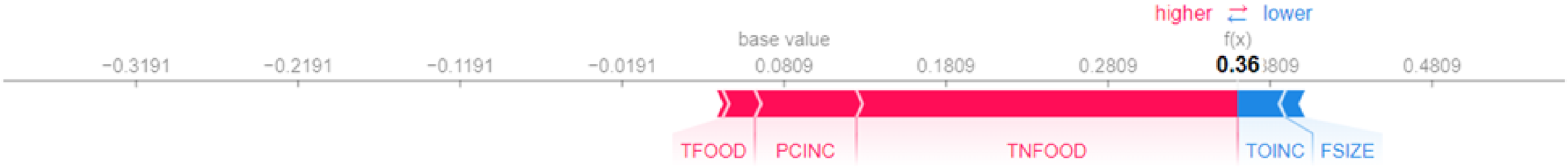


Results and Findings

Figure 4. Summary density scatter plot of SHAP values for each feature

Results and Findings

Figure 5. SHAP Force plot to interpret how the XGBoost model arrived at the prediction.



Conclusions

- This research trains and evaluates five binary classification models on the data and selects XGBoost as the best classifier—with an accuracy score of 0.67 and a ROC-AUC score of 0.72.
- For post-hoc analysis, utilizing SHAP values effectively visualized, identified, and interpreted the most important features influencing the model's predictions. In ascending order:
  - Top 3 features that influence the overall model: Total Non-Food Expenditure (TNFOOD), Per Capita Income (PCINC), and Total Income (TOINC).
  - Features that influence Urban(0): TNFOOD, PCINC, TFOOD, TTOTEX
  - Features that influence Rural(1): TOINC, FSIZE

Recommendations

- To improve food security and nutrition, policies should address food accessibility and affordability in both urban and rural areas.

- To achieve sustainable development, policies should consider the environmental and natural resource impacts of urbanization and rural preservation.

- Only six features were used for this classification research, which may not capture the full complexity of the variables.

Recommendations

- Other factors, such as geographic location, infrastructure and service access, social and cultural norms, and environmental conditions should also be considered.

- Further research such as exploring other machine learning techniques, incorporating spatial and temporal data, and validating the model with external data sources are beneficial and advantageous.