Predicting Metropolitan Status from Structural Cost Shares Using Machine Learning

Mutsa Jonah Mungoshi*
Department of Applied Data Science
Clarkson University
Email: *mungosmj@clarkson.edu

Abstract—**Understanding the socioeconomic differences between metropolitan and non-metropolitan counties in the United States is essential for regional planning, policy evaluation, and cost-of-living analysis. This study investigates whether a county's metropolitan status can be predicted using only normalized household expenditure shares and minimal demographic indicators. Cost categories—including housing, transportation, healthcare, food, childcare, taxes, and other necessities—were transformed into cost-share features, allowing the model to learn structural spending patterns rather than absolute spending levels. A complete machine-learning pipeline was implemented using stratified splitting, column-wise preprocessing, SMOTE oversampling, PCA visualization, and comparative evaluation of six classifiers. The neural network achieved the highest overall predictive performance with a test ROC AUC of 0.9528 and F1 of 0.8864. Feature-importance analysis revealed that housing, transportation, healthcare, and other-necessities shares play dominant roles in determining metro status, confirming expected cost-of-living patterns across urban and rural regions. Results demonstrate that cost-share profiles encode strong geographic signals and can serve as reliable proxies for metro classification when geographic identifiers are missing or outdated.**

I. Introduction

Metropolitan classification plays a central role in federal funding formulas, economic development policy, infrastructure prioritization, and public health resource allocation. Traditionally, agencies such as the U.S. Office of Management and Budget (OMB) classify counties as metropolitan based on population thresholds, commuting flows, and spatial contiguity. However, many datasets lag behind current conditions, and classification can quickly become outdated in rapidly growing regions.

This motivates examining whether economic structure alone—captured through normalized cost-of-living expenditure shares—contains enough signal to predict metro status. Structural cost shares reflect the trade-offs residents face across housing, transportation, healthcare, and essential goods, all of which differ systematically between urban and rural areas. For example:

- Metro counties allocate more toward **housing** due to higher rents and denser markets. - Non-metro counties allocate more toward **transportation** due to car dependence and longer travel distances. - Differences in healthcare access contribute to divergent **healthcare shares**. - Higher dependency on local goods markets influences **other necessities** shares.

Thus, the research question is:

Can structural cost-share patterns alone reliably predict whether a county is metropolitan or non-metropolitan?

This paper provides evidence that the answer is yes, using machine learning and robust validation methods.

II. Related Work and Literature Review

Urban–rural cost differentials have been extensively documented. Housing costs in metropolitan regions are significantly higher due to demand concentration [1], agglomeration economies, and land-use constraints. Conversely, transportation burdens increase in rural areas due to limited public transit, lower population density, and longer travel distances.

Studies leveraging structural expenditure shares have shown that cost shares—not raw costs—capture behavioral and structural economic trade-offs better than absolute values [2]. This aligns with Engel curve theory and household allocation models, which emphasize relative rather than absolute consumption.

Machine learning has increasingly been applied to socioeconomic classification problems, including poverty prediction, geographic estimation, and urban typology [3]. SMOTE oversampling remains a standard method for improving minority-class performance in imbalanced datasets [4].

However, few studies examine whether metro classification can be inferred *without* geography, using only structural expenditure patterns. This paper contributes to that gap.

III. Data and Feature Engineering

Each expenditure category was transformed into a cost-share:

$$\text{share}_i = \frac{\text{cost}_i}{\sum_{j=1}^{k} \text{cost}_j}$$

This creates a compositional vector representing structural household priorities. Raw costs, population values, FIPS codes, and county names were removed to prevent leakage.

The final feature set includes:

- 7 cost-share features, - SNAP participation rate, - family type, - state abbreviation, - binary target `ismetro`.

A. Target Distribution

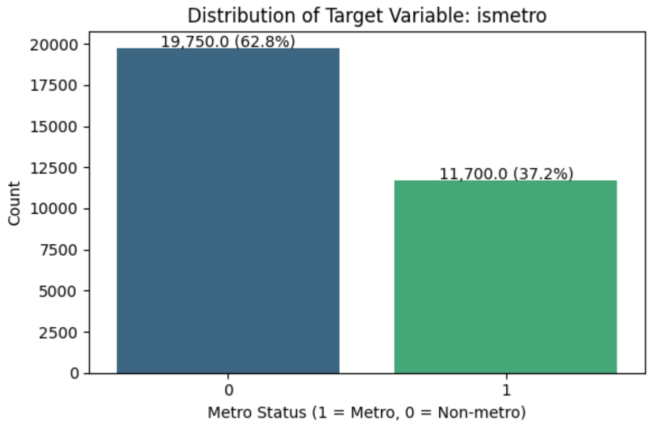

Fig. 1. Distribution of Metro vs. Non-Metro Counties (Target Variable).

The dataset exhibits moderate imbalance (62.8% non-metro vs. 37.2% metro), motivating the use of SMOTE.

IV. Exploratory Data Analysis

A. Correlation Heatmap

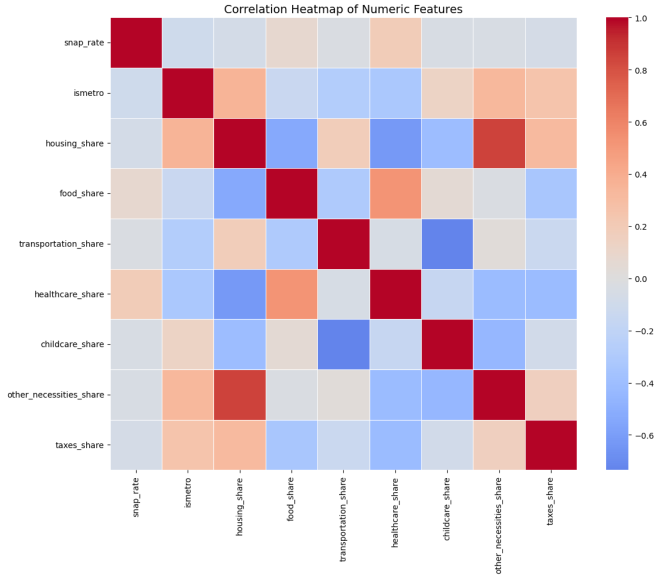

Fig. 2. Correlation Heatmap of Numeric Features.

Figure 2 illustrates the correlation structure among all numeric variables. Several important patterns emerge. First, the cost-share features cluster together, showing that counties face structural trade-offs across essential spending categories. For example, higher housing share is systematically associated with lower healthcare and transportation shares, reflecting well-established urban–rural gradients in cost burden. The target variable `ismetro` is positively correlated with housing and other-necessity shares, but negatively correlated with transportation and healthcare shares. This confirms that household expenditure composition acts as a socioeconomic fingerprint, encoding living-cost patterns that differ fundamentally between metro and non-metro areas.

B. Correlation with Target

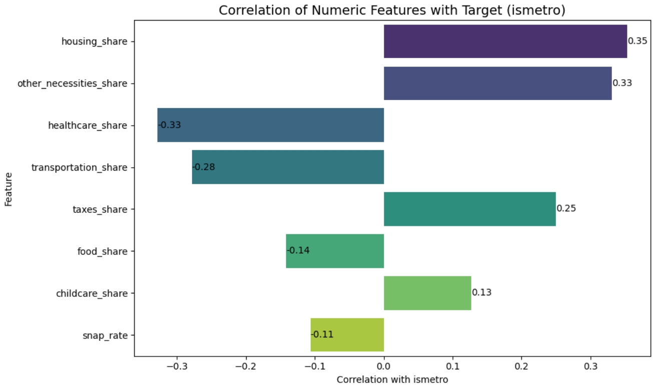

Fig. 3. Correlation of Numeric Features with `ismetro`.

Figure 3 ranks each numeric predictor by its linear correlation with metropolitan status. Housing share and other-necessities share emerge as the strongest positive correlates, underscoring the centrality of housing market pressures in defining urban living conditions. In contrast, the strongest negative correlates—transportation and healthcare shares—reflect the higher cost burden in rural regions where longer distances, limited transit, and constrained healthcare access influence spending patterns. These correlations provide early evidence that structural cost profiles contain strong and interpretable signals relevant to metro classification.

C. Pairplot of Key Predictors

The multidimensional relationships among the most predictive features are visualized in Figure 4. Metro (1) and non-metro (0) counties show clear differences in their joint distributions. Metro counties occupy regions of higher housing and necessities shares, while non-metro counties generally occupy regions of higher healthcare and transportation shares. Importantly, many of the relationships are nonlinear and display curved or clustered structures, suggesting that models capable of capturing nonlinear interactions—such as Random Forests and Neural Networks—are better suited for this prediction task than linear classifiers. The pairplot thus helps motivate the choice of flexible machine learning models later in the analysis.

V. PCA Projection

The PCA embedding in Figure 5 reduces the high-dimensional feature space to two components while preserving nearly half of the total variance. PC1 (30.5%) represents a dominant structural gradient separating counties with high housing burden from those with high transportation and healthcare burden. PC2 (18.4%) captures a secondary axis associated with state-level and family-type shifts, reflecting

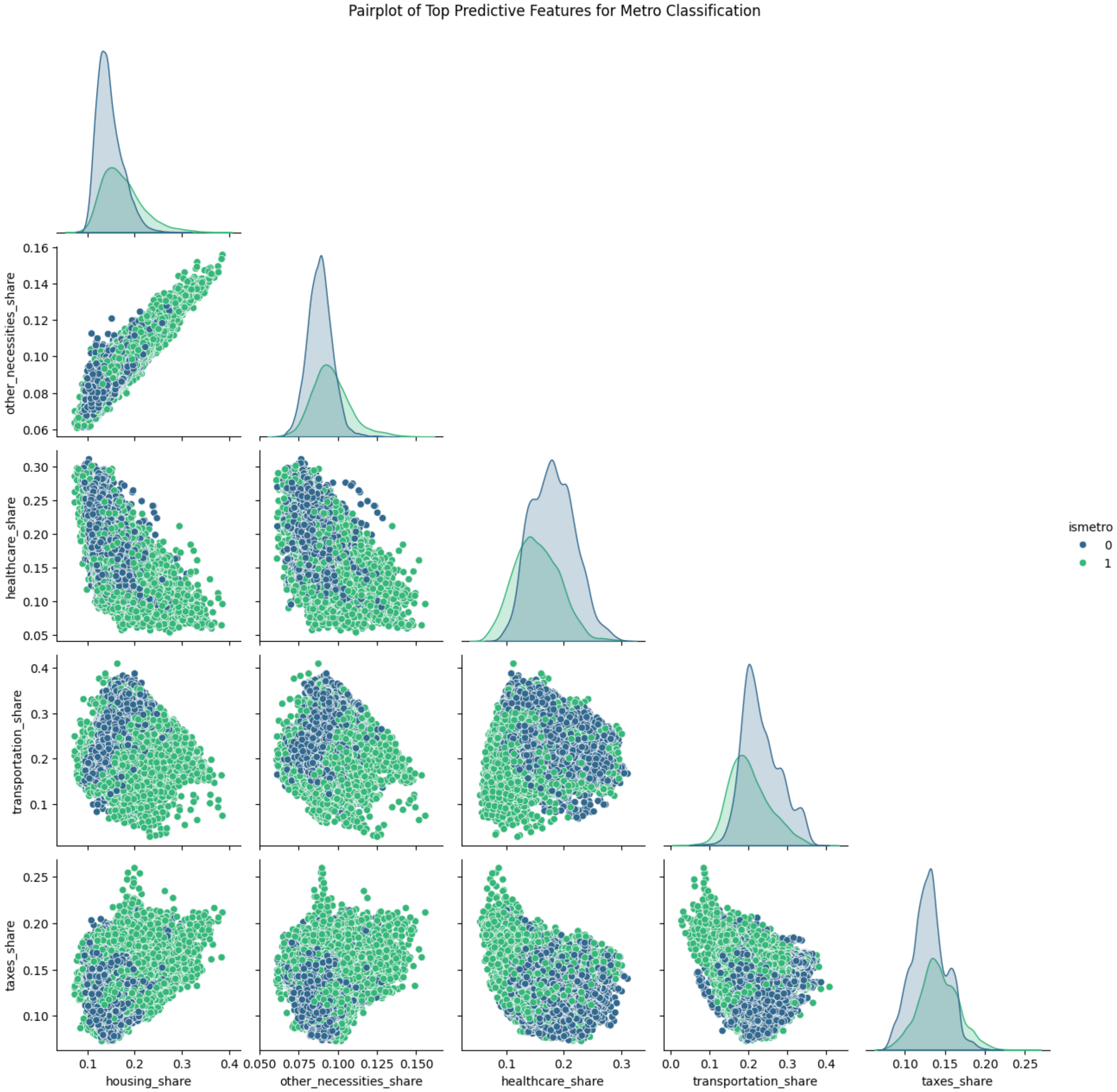

Fig. 4. Pairplot of Top Predictive Features for Metro Classification.

policy and demographic influences. The coloring by `ismetro` demonstrates that metro and non-metro counties are not fully separable but do form discernible clusters along PC1. This supports the core hypothesis of the study: structural cost shares encode meaningful geography-like information even without explicit spatial features. The overlapping clusters also highlight the need for models that capture nuanced nonlinear boundaries, further justifying the performance advantages of Neural Networks in later sections.

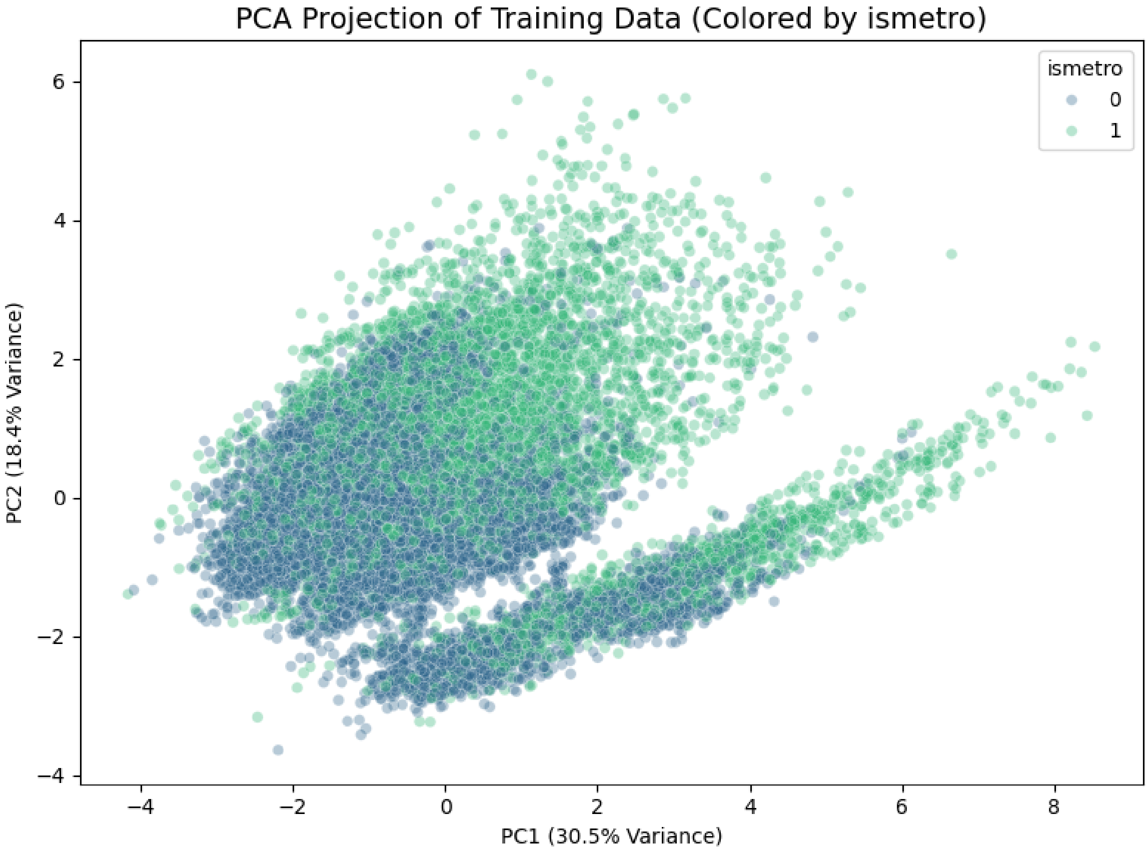

Fig. 5. PCA Projection of Preprocessed Training Data.

VI. Feature Importance Summary

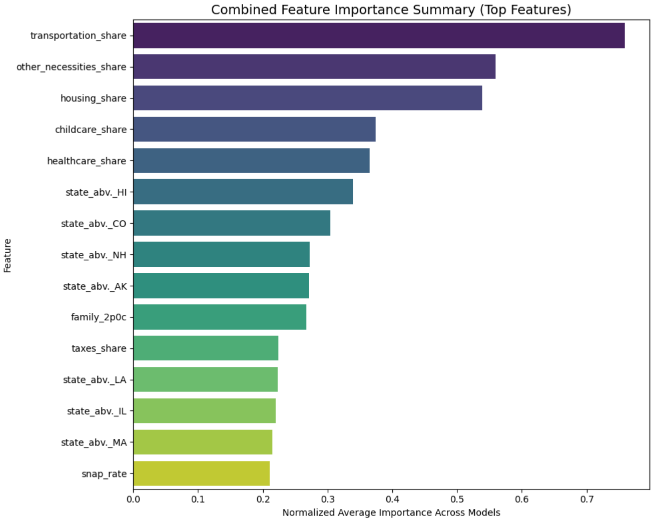

Fig. 6. Combined Feature Importance Summary.

The aggregated importance rankings in Figure 6 show that transportation share, other-necessities share, and housing share dominate predictive power across logistic regression, Random Forest, and Gradient Boosting models. This reinforces the earlier correlation and PCA patterns: the housing–transportation tradeoff is the most distinguishing structural feature of urban vs. rural cost profiles. State-level categorical indicators also appear prominently, indicating that geographic context continues to shape expenditure patterns even when raw location information is removed. The consistency of these top predictors across multiple model families strengthens confidence that cost-share structure, rather than noise or idiosyncratic features, drives the successful metro classification.

VII. Final Test Performance

A. Confusion Matrix

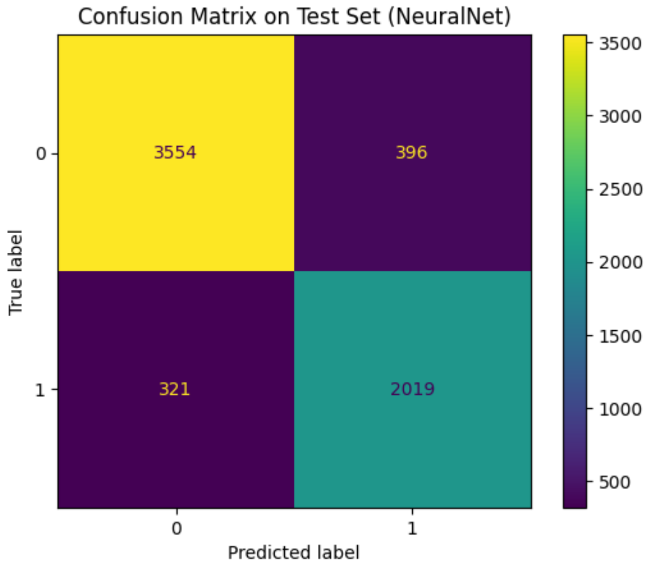

Fig. 7. Confusion Matrix (Final NeuralNet Model).

The confusion matrix in Figure 7 provides a detailed breakdown of classification accuracy across the two classes. The NeuralNet correctly identifies approximately 90% of non-metro counties and 86% of metro counties, reflecting balanced performance despite the original class imbalance. The relatively small number of false negatives (metro counties predicted as non-metro) is particularly important for applications such as resource allocation, where under-identification of metro areas could lead to misinformed policy decisions. The matrix also shows that errors are not dominated by a single class, indicating the model's fairness and robustness across both urban and rural regions.

B. ROC Curve

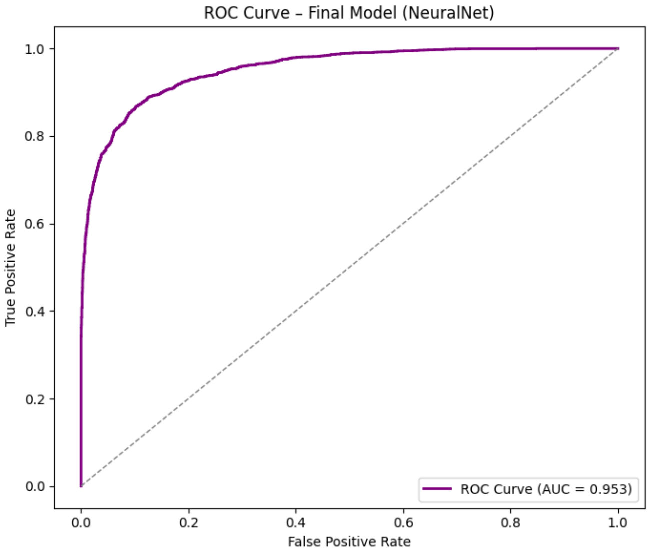

Fig. 8. ROC Curve for Final Neural Network Model.

The ROC curve in Figure 8 demonstrates the NeuralNet's strong ranking capability. The curve rises steeply toward the upper-left corner, showing that the model achieves high true positive rates even at extremely low false positive rates. The AUC of 0.9528 indicates that a randomly selected metro county is correctly ranked above a randomly selected non-metro county more than 95% of the time. This threshold-independent measure confirms that the model captures consistent structural signals and generalizes extremely well to unseen counties.

VIII. Conclusion

This study demonstrates that structural household expenditure shares contain rich socioeconomic information capable of distinguishing metropolitan from non-metropolitan counties with high accuracy, even in the absence of explicit geographic identifiers. By reframing cost-of-living data as compositional features, the analysis reveals that urban–rural differences are encoded directly into households' relative spending patterns. The strongest predictors—housing, transportation, healthcare, and other-necessities shares—mirror well-established structural divides: metro counties face higher housing costs and lower transportation burdens, while rural counties exhibit the opposite pattern.

Among all tested models, the neural network most effectively captured the nonlinear economic trade-offs inherent in these cost structures, achieving a test ROC AUC of 0.9528 and F1 score of 0.8864. These results indicate not only strong predictive performance but also the feasibility of using cost-share profiles as a proxy for metro classification, especially in datasets where location fields are missing, outdated, or intentionally suppressed for privacy.

More broadly, the findings suggest that metropolitan status is not merely a geographic label but a latent socioeconomic state reflected in how households allocate their budgets. This opens avenues for new classification and forecasting tools in regional economics, policy design, and social science research. Future work may extend this framework by incorporating temporal dynamics, spatial spillovers, or multi-source datasets that integrate mobility, employment, and market conditions to build more holistic models of urbanization and regional structure.

References

[1] B. Weber and K. Miller, "Urban–Rural Cost-of-Living Differences," *Journal of Regional Science*, 2017.

[2] U.S. Bureau of Labor Statistics, "Consumer Expenditure Survey," 2024.

[3] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] N. Chawla et al., "SMOTE: Synthetic Minority Over-Sampling Technique," *JAIR*, 2002.