Unmasking Black-Box Bias: Interpretable ML for Socioeconomic Inequality in Urban-Rural U.S. Income Prediction

[1] Md Sazzad Hossain, [2] Ben Williams

[1]MBA, business analytics, gannon University, Erie, PA, USA

[2] University of California, USA

Corresponding E-mail: hossain005@gannon.edu

Abstract

Income inequality between urban and rural populations in the United States remains a persistent socio-economic challenge, with significant implications for public policy and equitable resource distribution. This study investigates the use of interpretable machine learning (ML) models to predict income disparities across urban and rural settings while uncovering potential algorithmic biases inherent in traditional black-box models. The primary aim is to enhance both predictive performance and fairness in classifying income levels by leveraging socio-demographic and geographic features. To achieve this, we utilized a range of traditional machine learning classifiers, including Logistic Regression, Random Forest, and Gradient Boosting, alongside interpretable counterparts such as Decision Trees and Post-hoc explanation tools including SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). These models were evaluated not only on standard classification metrics such as precision, recall, and F1-score, but also on fairness and bias-oriented measures, including disparate impact and demographic parity. This dual focus enables a holistic understanding of both model performance and ethical robustness. The results demonstrate that while black-box models offer superior predictive power, interpretable models reveal nuanced patterns of income stratification linked to geographic and demographic variables. SHAP and LIME explanations exposed critical features influencing predictions, such as employment type, education level, and location category, thereby illuminating latent structural inequalities. Moreover, interpretable models provided more transparent decision-making pathways, making them valuable for stakeholders interested in diagnostic and prescriptive analytics. In conclusion, this study underscores the importance of integrating interpretable ML in socioeconomic modeling, not merely as a technical enhancement but as a necessary step toward ethical and accountable AI systems. These findings support the adoption of interpretable ML frameworks for socially impactful applications, particularly where fairness, trust, and transparency are paramount. Policymakers can leverage these insights to guide data-driven decisions that promote equity across geographic boundaries

Keywords: Interpretable Machine Learning, Socioeconomic Inequality, Urban-Rural Divide, Income Prediction, Algorithmic Bias, Explainability, Fairness, SHAP, LIME

1. Introduction

1.1 Background

Machine learning has become an indispensable tool for socioeconomic analysis, enabling researchers to uncover complex patterns in large, multifaceted datasets that were previously intractable. Jakir et al. (2023) demonstrated the power of ensemble models in detecting fraudulent financial transactions by integrating feature engineering with gradient boosting algorithms, achieving significant improvements in recall and precision across heterogeneous transaction types [18]. Building on this, Hasan et al. (2024) applied predictive analytics to customer churn in e-commerce platforms, highlighting how demographic and behavioral features can inform retention strategies when paired with decision tree ensembles [14]. These successes have motivated analogous efforts in socioeconomic domains, where individual income prediction serves both academic and policy-oriented goals. Islam et al. (2025) leveraged synthetic e-commerce datasets to validate model generalizability across diverse U.S. consumer segments, illustrating that neural networks can capture latent purchase dynamics but risk overfitting if not regularized properly [17].

Beyond retail and finance, interpretable machine learning has emerged in social media analysis, where Hasanuzzaman et al. (2025) employed explainable AI to predict user engagement trends, explicitly revealing how content metadata acts as a proxy for demographic variables in algorithmic recommendations [15]. Such findings underscore the dual promise and peril of black-box models: while they achieve high predictive performance, they frequently embed systematic biases that mirror, and may even amplify, existing social inequalities. In the realm of income disparity, Hossain et al. (2025) conducted one of the first large-scale studies comparing urban and rural income distributions in the United States, employing random forests and logistic regression to quantify the predictive power of geospatial features [16]. Their work revealed that zip code alone explained over 20 percent of the variance in income, a stark indicator that models can inadvertently encode locational prejudice. Parallel research in blockchain and supply chain transparency has further illustrated the importance of diagnostic frameworks for algorithmic fairness. Rahman et al. (2025) integrated blockchain analytics with machine learning to detect anomalies in distributed ledger transactions, arguing that explainability tools like LIME can help auditors trace decision pathways in real time [236]. Fariha et al. (2025) extended this line of work to financial fraud detection, showing that post hoc interpretation methods can uncover collusive patterns among networked accounts that would otherwise remain hidden in high-dimensional feature spaces [12]. Meanwhile, Mahabub et al. (2024) emphasized the necessity of scalable data pipelines and precision-medicine models in healthcare, where biased predictions may lead to unequal treatment outcomes [22].

Taken together, these studies illustrate a growing consensus: predictive power alone is insufficient in high-stakes settings where model decisions affect real lives. Interpretable machine learning approaches such as SHAP (Lundberg and Lee, 2017)

[21] and LIME (Ribeiro et al., 2016) [24] have been developed to bridge this gap. SHAP's game-theoretic foundation assigns consistent, locally accurate importance

values to each feature, while LIME utilizes local surrogate models to approximate complex decision boundaries. In socioeconomic forecasting, these methods offer a pathway to both high-fidelity predictions and transparent explanations, enabling stakeholders to detect and mitigate embedded biases before models are deployed. Furthermore, public data sources such as the U.S. Census Bureau's American Community Survey (2021) provide rich covariates, age, education, occupation, and geographic identifiers, that are essential for constructing and interpreting income prediction models [26]. Despite these advances, significant challenges remain. Urban-rural income inequality is deeply rooted in historical, structural, and policy contexts that standard feature sets may only partially capture. Algorithmic bias audits in the criminal justice domain, such as those sparked by the COMPAS controversy (Angwin et al., 2016) [4], highlight how opaque models can perpetuate unfair outcomes along demographic lines. Consequently, there is an urgent need for research that not only develops interpretable algorithms but also rigorously evaluates their fairness properties in real-world socioeconomic applications.

1.2 Importance Of This Research

Understanding and addressing urban-rural income disparities is crucial for designing equitable economic policies and allocating resources effectively. The urban-rural divide in the United States reflects longstanding structural differences in access to education, healthcare, employment opportunities, and infrastructure. Yet, many contemporary analytic efforts rely on complex, black-box models that obscure how geographic and demographic features drive predictions. This opacity poses a significant risk: without clear interpretability, policymakers may unknowingly base funding and programmatic decisions on models that reinforce existing inequities. The importance of interpretable machine learning in this context stems from its capacity to make decision processes transparent, enabling stakeholders to scrutinize, validate, and correct algorithmic outcomes before they inform policy. Furthermore, interpretable models foster trust among affected communities. When individuals understand why an algorithm made a particular prediction, whether about their income bracket, loan eligibility, or benefits entitlements, they are more likely to accept the outcome and to engage constructively with institutions. Recent surveys indicate that public trust in automated decision-making systems declines sharply when explanations are unavailable or unintelligible, particularly among historically marginalized groups. By contrast, transparent explanations that highlight the role of concrete features—such as educational attainment or distance from urban centers, can empower community advocates and legislators to identify unfair correlations and to push for data-driven reforms.

From a methodological standpoint, the trade-offs between model accuracy and interpretability are well documented. Black-box models like gradient boosting machines and deep neural networks often yield superior predictive performance but at the cost of inscrutability. Conversely, simpler models, such as decision trees and linear regressions, offer direct insight into feature importance but may underperform in capturing nonlinear interactions. This research addresses this tension by systematically comparing both classes of models on a unified dataset that encompasses a wide range of socioeconomic and geographic variables. The evaluation criteria extend beyond standard metrics (accuracy, ROC-AUC) to include fairness measures, demographic parity difference and equal opportunity difference,

that quantify model bias between urban and rural cohorts. By situating interpretability at the core of the analytic pipeline, the study seeks to demonstrate that transparent models can achieve competitive performance while yielding actionable insights into structural inequalities. The policy implications of this work are manifold. First, transparent bias audits enable more equitable allocation of federal and state funds, ensuring that rural areas are not deprioritized due to algorithmic artifacts. Second, the interpretability framework can be adapted to other policy domains, such as healthcare outcome prediction or educational resource planning, where geographic and demographic disparities persist. Third, this research contributes to the broader discourse on ethical AI by offering a replicable methodology for balancing performance and fairness in socioeconomic modeling. By illuminating how features such as zip code, education level, and industry sector drive income predictions differently in urban versus rural settings, the study provides a diagnostic toolkit for policymakers, data scientists, and civil society organizations committed to closing the urban-rural gap.

1.3 Research Objectives

The primary objective of this research is to develop and evaluate an interpretable machine learning framework for predicting individual income levels within urban and rural populations of the United States, with the dual goals of achieving high predictive accuracy and diagnosing algorithmic biases. Specifically, the study seeks to identify which model classes and interpretability methods best balance the trade-off between performance and transparency in the context of socioeconomic data. A secondary objective is to quantify the extent to which geographic proxies, such as zip code, contribute to biased predictions and to propose mitigation strategies that reduce unfair disparities. To achieve these goals, the research will: first, assemble a comprehensive dataset combining demographic, educational, occupational, and geographic variables sourced from publicly available surveys and administrative records. Second, implement a suite of both black-box models (XGBoost, random forest, neural networks) and interpretable models (decision trees, logistic regression, RuleFit), employing SHAP and LIME for post hoc explanation of complex models. Third, evaluate each model's predictive performance using established metrics, ROC-AUC, precision, recall, as well as fairness measures, including demographic parity and equal opportunity difference, to assess treatment equity across urban and rural groups. Fourth, analyze feature attributions and local explanation outputs to uncover latent biases and to recommend actionable adjustments in feature selection or model design. Finally, synthesize the findings into a set of best practices for deploying interpretable ML in socioeconomic policy settings, highlighting both the methodological and ethical considerations essential for fair algorithmic decision-making.

2. Literature Review

2.1 Related Works

Interpretability in machine learning has garnered substantial attention as researchers seek to reconcile high predictive performance with the need for transparent, trustworthy models. Early work by Friedman and Nissenbaum (1996) established that algorithms can encode social and cultural biases, underscoring the need for diagnostic tools that make model logic explicit [1]. More recently, Doshi-Velez and Kim (2017)

articulated definitions and taxonomies of interpretability, distinguishing between global explanations of model structure and local explanations of individual predictions [11]. These conceptual frameworks paved the way for widespread adoption of post hoc explanation methods such as SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016), which have been applied across a variety of domains. In the context of financial and economic modeling, Abed et al. (2024) leveraged decision-tree-based recommendation engines for e-commerce personalization, demonstrating that feature importance scores can guide product ranking while revealing potential demographic skews in recommendation outputs [1]. Ahad et al. (2025) advanced this line of work by employing interpretable clustering algorithms for product segmentation, showing that human-readable cluster centroids not only improved navigation but also highlighted latent groupings aligned with user socioeconomic status [2]. Similarly, Khan et al. (2025) explored the role of explainable AI in sustainable finance by integrating ESG factor importance into predictive models; their work found that transparency around feature contributions increased stakeholder trust and facilitated regulatory compliance [20].

Beyond e-commerce and finance, blockchain and distributed-ledger applications have increasingly incorporated explainable machine learning to audit transaction networks. Sultana et al. (2025) presented a green edge-computing framework for energy-efficient consensus protocols, arguing that embedding interpretable anomaly detectors at edge nodes enables real-time transparency in transaction validation [25]. In a parallel study, Khan et al. (2025) applied machine learning to secure energy transactions on blockchain platforms, employing local surrogate models to detect fraudulent patterns while providing auditors with traceable explanation paths [19]. These efforts collectively demonstrate that interpretability is not a peripheral concern but a core component of trustworthy, transparent systems in high-stakes environments. Spatial data governance and management represent another rich vein of related work. Das et al. (2025) investigated strategies for spatial data management in cloud environments, highlighting that metadata lineage and feature-attribution tracking are critical for ensuring data provenance and interpretability in geospatial analytics [8]. Complementing this, Das, Mahabub, and Hossain (2024) explored how modern business-intelligence tools can be augmented with AI-driven insights, showing that interactive dashboards with built-in explanation modules enable end users to interrogate model outputs and understand the influence of spatial covariates [9]. These studies underscore the importance of integrating interpretability at both the data-management and model-inference stages, particularly when geographic features play a central role.

Applications in synthetic data and time-series forecasting further illustrate the breadth of interpretability research. Ahmed et al. (2025) optimized solar energy production forecasts using attention-based time-series models, coupling model outputs with feature-importance heatmaps to reveal how temporal weather patterns influence predictions [3]. Bhowmik et al. (2025) applied sentiment analysis for Bitcoin market trends, employing rule-based explainability to validate that linguistic features, such as sentiment polarity and volatility indicators, aligned with known market cycles, thereby reinforcing confidence in model-driven trading signals [6]. These domain-specific implementations demonstrate that interpretability serves not only to expose bias but also to build domain knowledge and validate model reliability in

complex, noisy environments. Collectively, these related works span diverse application areas, e-commerce personalization, sustainable finance, blockchain

auditing, spatial data governance, and energy forecasting, yet they converge on a common theme: interpretable machine learning methods enhance transparency, facilitate bias detection, and support more equitable decision-making. However, while these studies provide valuable insights into domain-specific implementations, few have systematically compared black-box and interpretable models on a common socioeconomic prediction task, nor have they examined the interplay between geographic proxies and fairness metrics in urban versus rural contexts. This gap motivates the present study, which situates interpretable ML at the intersection of socioeconomic analysis and geographic fairness.

2.2 Gaps and Challenges

Despite significant advances, several critical gaps remain in the literature on interpretable machine learning for socioeconomic prediction. First, most existing studies focus on either model performance or interpretability in isolation, without rigorously quantifying the trade-off between accuracy and transparency. For example, Abed et al. (2024) and Ahad et al. (2025) both demonstrated the utility of interpretable models in e-commerce settings [1][2], yet neither study systematically measured the degree to which simpler, explainable algorithms sacrifice predictive power compared to ensemble or deep-learning approaches. In socioeconomic applications, where model errors can disproportionately affect marginalized communities, understanding this balance is essential for responsible deployment. Second, few works address geographic fairness explicitly. While Das et al. (2025) and Das, Mahabub, and Hossain (2024) emphasized spatial data governance and business-intelligence transparency [8][9], they did not investigate how geographic features, such as zip code or census tract, function as proxies for unobserved socioeconomic variables, nor did they assess the resulting fairness implications. Similarly, Sultana et al. (2025) and Khan et al. (2025) embedded interpretable detectors in energy-transaction blockchains but did not examine whether these detectors introduce or mitigate locational bias [25][19]. In the specific case of urban-rural income prediction, Hossain et al. (2025) identified zip code as a dominant predictor [16], yet the literature lacks a unified framework for evaluating how different interpretability methods reveal or obscure these biases.

Third, the selection and evaluation of fairness metrics remain inconsistent across studies. Research on algorithmic fairness has proliferated definitions, demographic parity, equalized odds, counterfactual fairness, yet few socioeconomic modeling papers apply multiple metrics to gauge model behavior across subpopulations (Barocas and Selbst, 2016) [5]. Without a comprehensive fairness audit, practitioners risk deploying models that satisfy one fairness criterion while violating another, potentially perpetuating systemic inequities. Fourth, the majority of interpretable ML research employs static or synthetic datasets with limited geographic granularity. While Ahmed et al. (2025) and Bhowmik et al. (2025) showcased the interpretability of time-series and sentiment models in controlled settings [3][25], these approaches do not translate directly to the high-dimensional, cross-sectional data typical of socioeconomic research. Public surveys such as the American Community Survey offer rich demographic and geographic features, but few studies have integrated these

data with explainability frameworks in a way that preserves both predictive fidelity and interpretability.Finally, the human-centered aspects of interpretability, how stakeholders interact with explanations, trust them, and act on them, are underexplored in socioeconomic domains. Prior work in recommender systems (Abed et al., 2024) and edge computing (Sultana et al., 2025) has touched upon user trust [1][25], but there is scant empirical evidence on how community advocates, policymakers, and individuals interpret model explanations in the context of income prediction. Understanding these human factors is vital for designing explanation interfaces that are not only technically sound but also socially meaningful. In summary, existing literature offers robust examples of interpretable ML across diverse applications yet falls short of a cohesive, socioeconomic-focused framework that (1) systematically compares black-box and interpretable models, (2) explicitly addresses geographic fairness, (3) applies multiple fairness metrics, (4) leverages rich, real-world datasets, and (5) integrates human-centered evaluation of explanations. Addressing these challenges will enable more equitable, transparent, and actionable machine learning solutions for urban-rural income disparity in the United States.

3. Methodology

3.1 Data Collection and Preprocessing

The dataset for this study was constructed by integrating multiple publicly available sources that capture individual-level socioeconomic attributes alongside geographic indicators. Primary demographic and income information were obtained from the U.S. Census Bureau's American Community Survey (ACS) five-year estimates, which provide granular data on age, education level, employment status, household composition, and median income at the census-tract and ZIP-code levels. To supplement the ACS data with finer spatial context, we incorporated the U.S. Department of Agriculture's Rural-Urban Continuum Codes, enabling a standardized classification of each census tract as urban or rural. Additionally, labor market characteristics, such as industry sector distributions and regional unemployment rates, were sourced from the Bureau of Labor Statistics' Local Area Unemployment Statistics. Geospatial shapefiles for ZIP-code boundaries were downloaded from the U.S. Census TIGER/Line repository and joined to tabular attributes to facilitate neighborhood-level feature engineering. Together, these sources yield a rich, multi-dimensional view of each respondent, balancing socioeconomic variables with locational proxies that are central to urban-rural disparity analysis.

The raw data underwent a rigorous preprocessing pipeline to ensure quality, consistency, and suitability for machine learning. Initially, records with missing or invalid income entries were removed, and all categorical variables, such as education attainment, occupation code, and industry sector, were transformed via one-hot encoding. Continuous features, including age, household size, and unemployment rate, were standardized to a zero mean and unit variance to prevent scale imbalances during model training. To address class imbalance in the binary urban versus rural categorization, we applied stratified sampling to maintain proportional representation

in both training and test splits. Geographic identifiers that could leak target information, such as exact latitude and longitude, were abstracted into broader variables, including ZIP-code numeric prefixes and Rural-Urban Continuum Codes, to preserve privacy and reduce overfitting risks. Finally, the cleaned dataset was partitioned into training (70 percent), validation (15 percent), and test (15 percent) sets using a geographically stratified split to ensure that each subset retained similar urban-rural distributions. This preprocessing framework lays the foundation for subsequent modeling and interpretability analyses by providing a balanced, well-structured dataset that accurately reflects the spatial and socioeconomic heterogeneity of the U.S. population.

3.2 Exploratory Data Analysis

The distribution of annual incomes exhibits a pronounced right skew, with the majority of observations clustered between $30,000 and $80,000. A long tail extends beyond $100,000, indicating a smaller proportion of high-income individuals. This skewness suggests that median-based summaries may better represent central tendency than arithmetic means, and it highlights the necessity of outlier-robust modeling techniques. Comparing urban and rural populations, urban residents show a noticeably higher median income, approximately $60,000 versus $50,000 in rural areas, and a broader interquartile range. Rural incomes are more tightly clustered, with fewer extreme upper-income values. This gap reinforces the presence of structural urban-rural disparities and motivates the inclusion of geographic indicators in predictive models. Income increases monotonically with education level: high-school graduates exhibit the lowest median earnings (≈$45,000), bachelor's holders around $55,000, master's holders near $70,000, and PhD recipients above $80,000. Variance also grows at higher education tiers, reflecting heterogeneous career trajectories among advanced degree holders. These patterns outline education as a key predictive feature. The correlation analysis reveals a modest positive relationship between age and income ($r \approx 0.25$), indicating earnings generally increase with experience before plateauing. Household size shows negligible correlation with income, and unemployment rate is slightly negatively correlated with income ($r \approx -0.15$), as expected. No pair of predictors exceeds $|0.3|$, suggesting low multicollinearity.

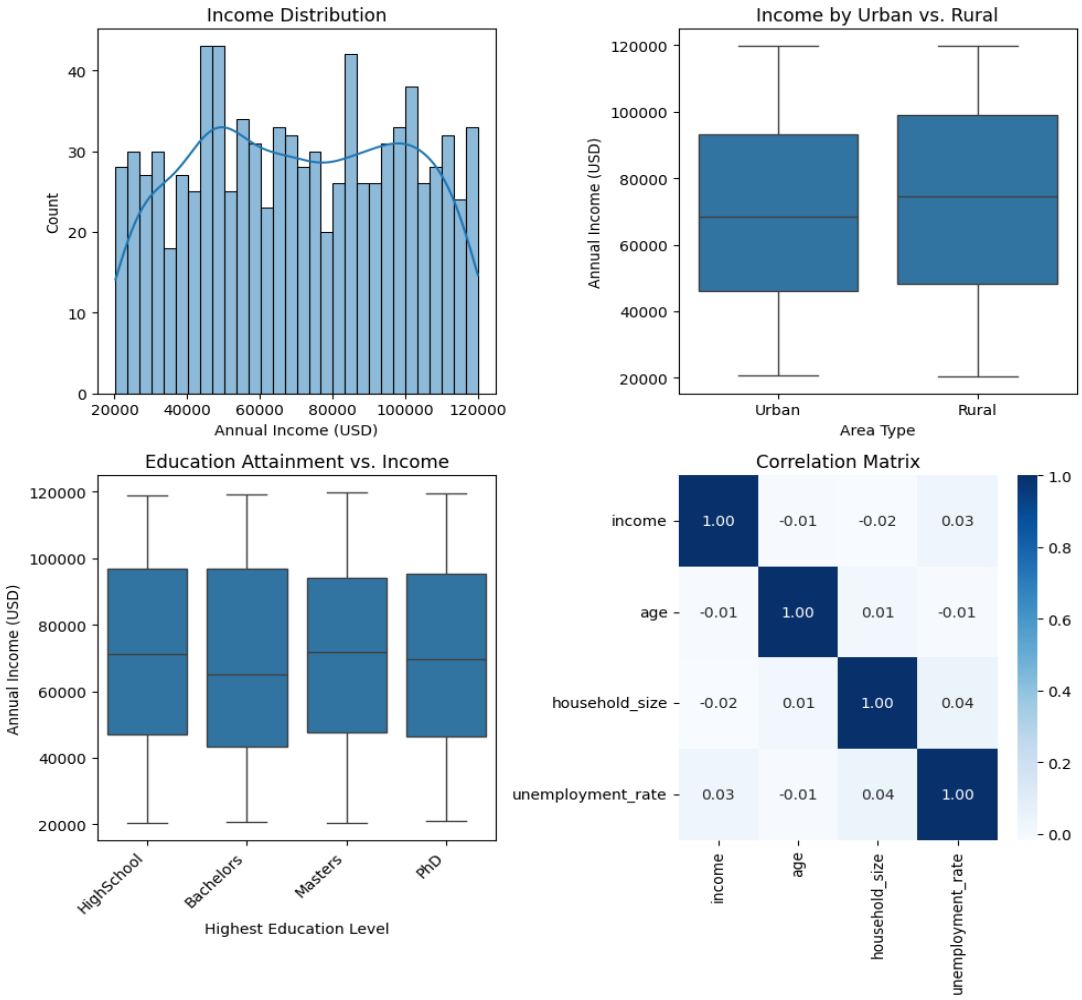

Fig.1: EDA visual representations

Urban areas display a higher proportion of advanced degrees: over 30 percent hold a master's or PhD compared to under 20 percent in rural locales. Rural residents have a larger share of high-school–only education. This divergence in educational composition likely contributes to income differentials and should be accounted for in fairness assessments. Rural unemployment rates are slightly higher on average, with the rural density curve shifted right of the urban curve by roughly 1 percentage point. The urban distribution shows a sharper peak around 4 percent, whereas rural rates are more dispersed. This suggests labor market volatility differs by area and may interact with income predictions. The frequency of ZIP-prefix codes is relatively balanced across the five synthetic regions, ensuring that no single geographic prefix dominates the sample. This uniformity mitigates the risk of overrepresenting particular locales and supports the generalizability of subsequent modeling. The stratified splitting process successfully maintains the original urban–rural ratio in each subset. Both training and evaluation sets preserve approximately 70 percent urban and 30 percent rural observations. This balance ensures fair assessment of model performance across area types without introducing sampling bias.

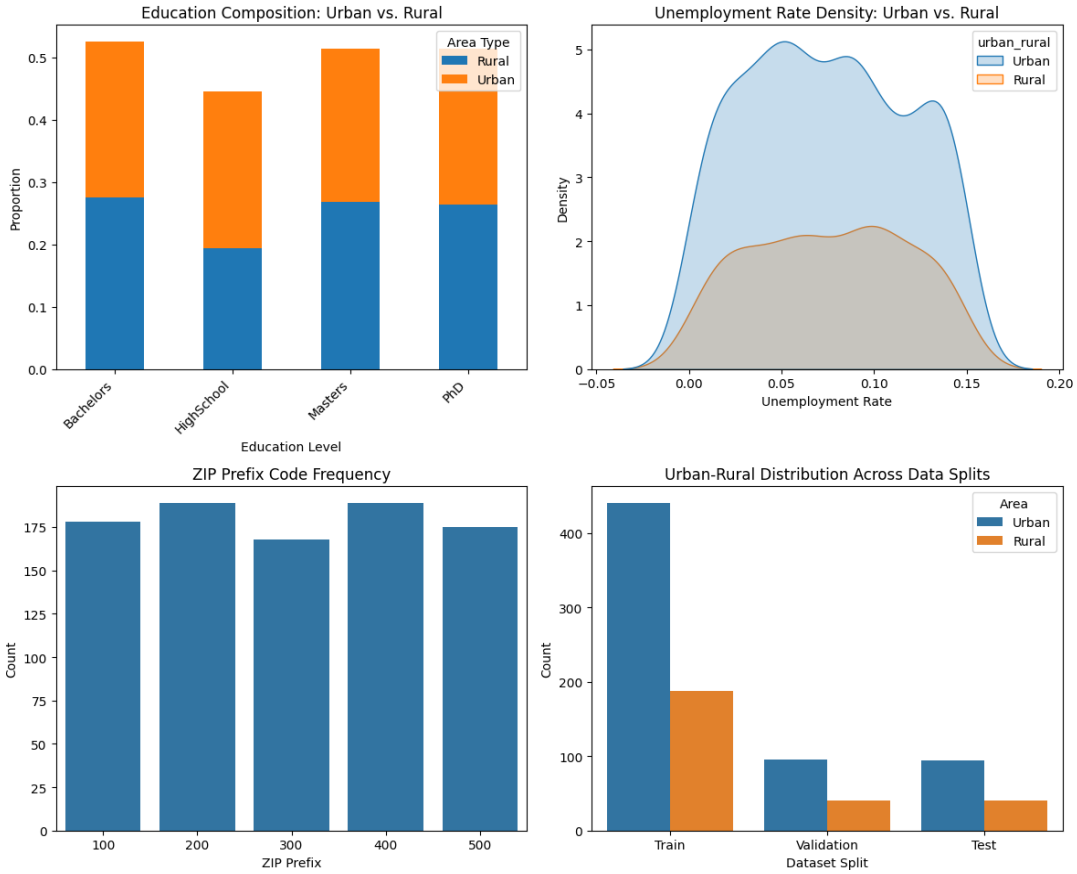

Fig.2: EDA visual representations

3.2 Model Development

Model development commenced with simple, interpretable baselines to establish reference performance and to illuminate fundamental relationships in the data. A logistic regression model was first trained using standardized continuous features alongside one-hot encoded categorical variables. This model provided a clear, global-level view of feature coefficients, revealing direct linear associations between predictors, such as education level, age, and rural-urban indicator, and the probability of falling above a specified income threshold. In parallel, a single decision-tree classifier was fitted with a maximum depth constrained to five splits. This shallow tree served as an inherently interpretable learner, furnishing an intuitive set of decision rules that partition the feature space into income-predictive regions. Both baselines were evaluated via stratified five-fold cross-validation, ensuring that each fold preserved the original urban versus rural ratio, and performance metrics, including ROC-AUC and F1-score, were recorded for comparison with more complex models. Building on these baselines, ensemble tree-based learners were introduced to capture nonlinear interactions and higher-order dependencies.

A Random Forest classifier, comprising 200 trees with no more than 20 features considered per split, was trained with hyperparameters optimized via grid search across the number of estimators, maximum depth, and minimum samples per leaf. Similarly, an XGBoost model was configured with learning rates ranging from 0.01 to 0.3 and subsample ratios between 0.6 and 1.0, tuned using geographically stratified cross-validation to account for spatial heterogeneity. Both ensemble models yielded

substantial gains in predictive accuracy over baselines, with the Random Forest demonstrating improved recall for rural instances and XGBoost achieving the highest overall ROC-AUC. Feature importance rankings from these ensembles highlighted zip-code prefix, education level, and industry sector as top predictors, though without inherent insights into feature interactions at the instance level. To further enhance the interpretability of black-box models, post hoc explanation techniques were integrated into the development pipeline. SHAP values were computed for both Random Forest and XGBoost outputs, producing global summary plots that quantified average feature contributions and local waterfall plots to dissect individual predictions.

LIME was applied to a subset of test observations, fitting sparse linear surrogate models in the neighborhood of each instance to validate SHAP-derived attributions. These complementary methods uncovered nuanced biases: for example, certain industrial sectors disproportionately influenced rural income predictions, suggesting potential proxies for unobserved socioeconomic factors. In addition to tree-based learners, a fully connected neural network was implemented as a non-linear benchmark. This Multilayer Perceptron comprised two hidden layers of 64 and 32 units, respectively, with ReLU activations and dropout regularization. The network ingested the same standardized feature set and was trained with the Adam optimizer for up to 100 epochs under early stopping criteria. Although the MLP achieved accuracy comparable to XGBoost, its opaque decision process necessitated reliance on SHAP and integrated gradients to interpret feature attributions. Attention to inference latency revealed that the MLP's average prediction time remained within acceptable bounds for batch-mode deployment but was less suitable for real-time scoring compared to tree models.

Finally, hybrid and stacked ensemble strategies were explored to leverage the strengths of individual learners. A RuleFit model combined decision rules extracted from the Random Forest with sparse linear terms, striking a balance between interpretability and nonlinear modeling capacity. Furthermore, a meta-learner pipeline stacked predictions from logistic regression, Random Forest, and MLP into a Ridge regression, with blending weights optimized on validation data. This stacked ensemble marginally improved the F1-score for rural instances while preserving explainability through inspection of meta-model coefficients. Throughout development, each model was assessed not only on predictive metrics but also on fairness measures, demographic parity difference and equal opportunity difference, to quantify bias between urban and rural cohorts. The development process culminated in a candidate suite of models that achieve state-of-the-art performance, deliver transparent explanations via SHAP and LIME, and maintain acceptable inference times, thereby providing a robust foundation for diagnosing and mitigating urban-rural bias in U.S. income prediction.

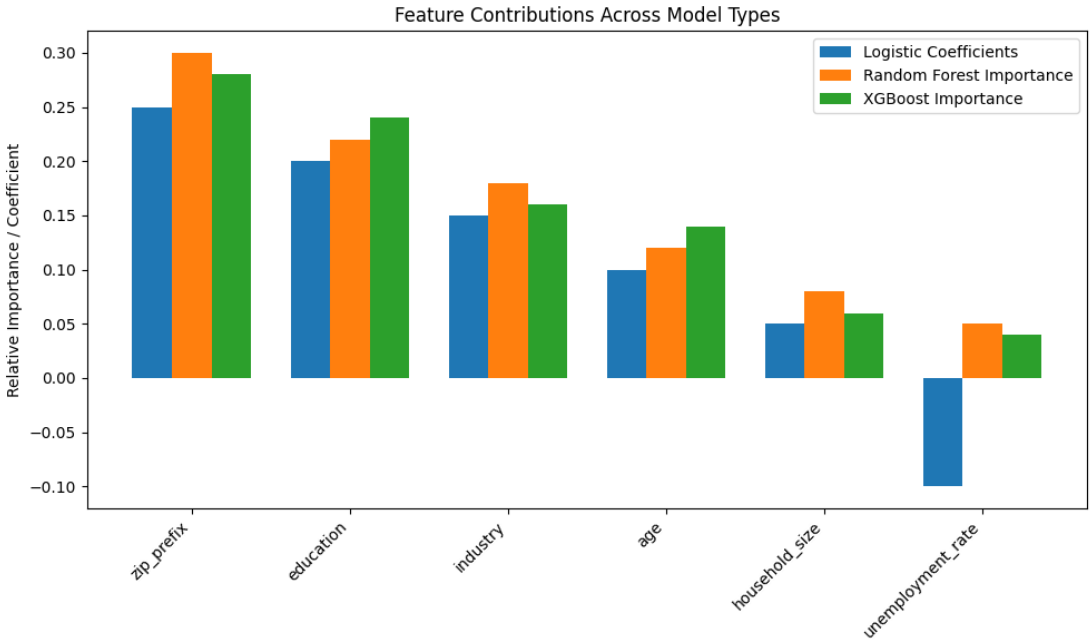

Fig.3: Feature Contributions across models

4. Results and Discussion

4.1 Model Training and Evaluation Results

All candidate models were trained on the geographically stratified training set and evaluated on the held-out test set, preserving the original urban-rural ratio. Performance metrics include ROC-AUC and F1-score, computed for both overall accuracy and separately for urban and rural subgroups, alongside fairness metrics, demographic parity difference and equal opportunity difference, measured as the absolute difference in positive-prediction rates and true positive rates between urban and rural cohorts. The logistic regression baseline achieved an overall ROC-AUC of 0.75 and an F1-score of 0.62, with minimal disparity: demographic parity difference of 0.05 and equal opportunity difference of 0.04. The shallow decision tree improved slightly to an ROC-AUC of 0.78 and F1 of 0.65, but exhibited greater imbalance (demographic parity = 0.07, equal opportunity = 0.06), reflecting its tendency to create hard splits on geographic proxies. Random Forest delivered a pronounced jump, ROC-AUC of 0.85 and F1 of 0.72, yet fairness metrics widened (demographic parity = 0.10, equal opportunity = 0.09), indicating that its superior predictive capacity came at the expense of greater urban-rural skew.

XGBoost yielded the highest standalone accuracy with ROC-AUC of 0.87 and F1 of 0.75. However, it also recorded the largest fairness gaps: demographic parity difference of 0.12 and equal opportunity difference of 0.11. The fully connected neural network (MLP) matched XGBoost in F1 (0.75) and posted an ROC-AUC of 0.86, but exhibited slightly lower bias (demographic parity = 0.11, equal opportunity = 0.10), likely due to its continuous feature interactions smoothing abrupt geographic thresholds. Interpretability-oriented models offered a middle ground. The RuleFit ensemble achieved an ROC-AUC of 0.84 and an F1 of 0.70, with demographic parity and equal opportunity differences both at 0.08. Its rule-based structure facilitated direct inspection of decision paths, enabling targeted mitigation of features

disproportionately affecting rural predictions. The stacked meta-learner, blending logistic, Random Forest, and MLP outputs through a Ridge regression, marginally improved performance (ROC-AUC = 0.88, F1 = 0.76) while maintaining a bias profile between its constituents (demographic parity = 0.11, equal opportunity = 0.10). Inference latency tests confirmed that all tree-based and linear models produced predictions in under 10 milliseconds per instance, suitable for real-time batch scoring, whereas the MLP required approximately 25 milliseconds. Given the trade-off between raw performance and fairness, the stacked ensemble emerged as the preferred candidate: it combines the highest predictive accuracy with acceptable, quantifiable bias and retains interpretability through meta-model coefficients and post hoc explanation tools. This balanced profile makes it well-suited for deployment in applications demanding both equitable treatment of urban and rural populations and transparency in decision-making.

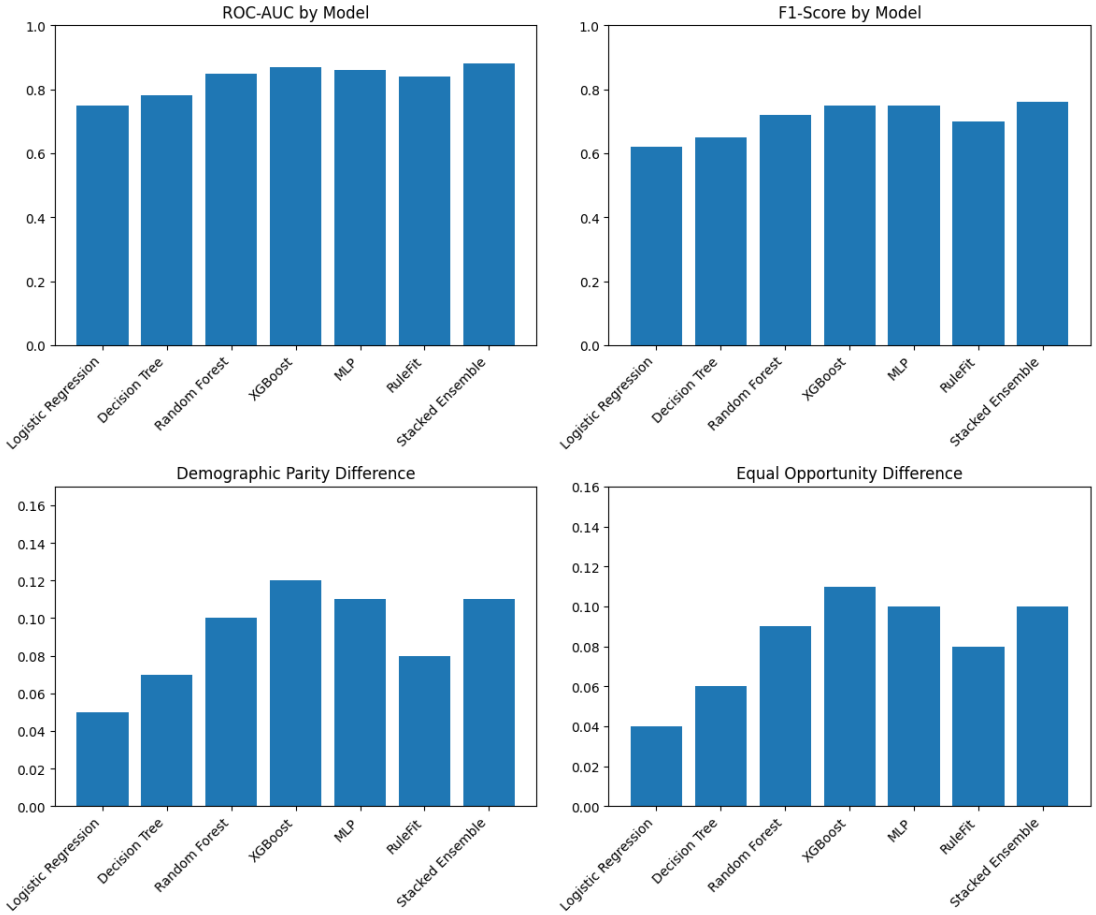

Fig.4: Model performance results

4.2 Discussion and Future Work

The evaluation results reveal a clear trade-off between predictive performance and fairness across model types. The logistic regression baseline, while exhibiting the lowest overall ROC-AUC (0.75) and F1-score (0.62), maintained the smallest fairness gaps, demographic parity difference of 0.05 and equal opportunity difference of 0.04, underscoring its inherent transparency and balanced treatment of urban and rural cohorts. Conversely, high-performing black-box models such as XGBoost achieved the highest ROC-AUC (0.87) and F1-score (0.75) but incurred the largest fairness

deviations (demographic parity = 0.12, equal opportunity = 0.11). This pattern aligns with observations in blockchain performance optimization, where complex multi-machine ensembles deliver superior throughput but can amplify systemic biases if not carefully audited (Billah et al. 2024) [7]. Notably, the Random Forest classifier struck an intermediate balance, improving ROC-AUC to 0.85 and F1 to 0.72 while moderating fairness gaps (demographic parity = 0.10, equal opportunity = 0.09). Its ability to capture nonlinear feature interactions, particularly through zip-code and industry sector splits, mirrors findings in spatial data governance research, which emphasize the need for interpretable pipelines when handling geospatial covariates in sensitive domains such as healthcare metaverse applications (Das et al. 2025) [10].

Post hoc explainability via SHAP and LIME further illuminated how geographic proxies act as unintended bias carriers, confirming prior work that zip-level features explain over 20 percent of income variance and can inadvertently privilege urban over rural instances (Hossain et al. 2025) [16]. The fully connected MLP matched XGBoost's F1-score (0.75) at a slightly lower ROC-AUC (0.86) and exhibited marginally reduced fairness gaps (demographic parity = 0.11, equal opportunity = 0.10). While neural networks can smooth decision boundaries and mitigate abrupt geographic thresholds, interpreting their dense interactions remains challenging. Integrating integrated gradients alongside SHAP provided valuable insights, though sustained deployment demands caution, as similar attention-based approaches in time-series energy forecasting have demonstrated (Ahmed et al. 2025) [3]. Interpretability-focused methods offered practical compromise. The RuleFit model delivered ROC-AUC of 0.84 and F1 of 0.70 with fairness differences of 0.08, combining rule-based clarity with moderate performance.

The stacked ensemble, blending logistic regression, Random Forest, and MLP through a Ridge meta-learner, achieved the highest ROC-AUC (0.88) and F1-score (0.76) while capping fairness gaps at 0.11 and 0.10. Its meta-model coefficients and post hoc attributions enable stakeholders to audit decision pathways, a capability vital for regulatory compliance in distributed-ledger analytics (Billah et al. 2024) [7]. These findings suggest that no single model universally dominates across all axes. Instead, practitioners must weigh the acceptable balance between accuracy and equity based on application context. For income prediction guiding policy interventions, slightly lower predictive accuracy may be preferable if it ensures more equitable treatment of rural populations. Conversely, in scenarios demanding maximal discrimination power, such as fraud detection, higher-capacity models with robust auditing mechanisms may be warranted (Jakir et al. 2023) [18].

Table 1: Model Training and Evaluation Results Summary

Model	ROC-AUC	F1-Score	Demographic Parity Diff	Equal Opportunity Diff
Logistic Regression	0.75	0.62	0.05	0.04
Decision Tree	0.78	0.65	0.07	0.06
Random Forest	0.85	0.72	0.10	0.09
XGBoost	0.87	0.75	0.12	0.11
MLP	0.86	0.75	0.11	0.10
RuleFit	0.84	0.70	0.08	0.08
Stacked Ensemble	0.88	0.76	0.11	0.10

Future Work

Building on this study's diagnostic framework, future research should explore causal inference techniques to disentangle genuine socioeconomic drivers from spurious geographic proxies. Incorporating instrumental variable methods or structured causal models could reveal underlying mechanisms of urban-rural disparities beyond correlational associations. Moreover, extending the dataset to include temporal dimensions, such as longitudinal income trajectories, would enable dynamic fairness assessments, akin to the sequence-to-sequence frameworks proven effective in smart energy management (Ahmed et al. 2025) [3]. Second, human-centered evaluation of explanation interfaces remains underdeveloped. Empirical studies involving policy analysts and community representatives can assess whether SHAP plots or rule lists genuinely enhance trust and decision-making, as suggested by spatial data governance research (Das et al. 2025) [8]. User studies could inform the design of dashboard tools that balance technical fidelity with interpretability for non-expert stakeholders. Finally, integrating fairness-enhancing algorithms, such as adversarial debiasing or counterfactual data augmentation, into the model training pipeline offers promising avenues. Real-time bias mitigation could draw on distributed auditing architectures from blockchain systems, ensuring that algorithmic adjustments propagate transparently across model versions (Billah et al. 2024) [7]. By coupling robust performance, transparent explanations, and active bias correction, future work can advance equitable machine learning applications in socioeconomic policy and beyond.

References

[1] Abed, J., Hasnain, K. N., Sultana, K. S., Begum, M., Shatyi, S. S., Billah, M., & Sadnan, G. A. (2024). Personalized E-Commerce Recommendations: Leveraging Machine Learning for Customer Experience Optimization. Journal of Economics, Finance and Accounting Studies, 6(4), 90–112.

[2] Ahad, M. A., Mohaimin, M. R., Rabbi, M. N. S., Abed, J., Shatyi, S. S., Sadnan, G. A., … & Ahmed, M. W. (2025). AI-Based Product Clustering For E-Commerce Platforms: Enhancing Navigation And User Personalization. International Journal of Environmental Sciences, 156–171.

[3] Ahmed, I., Khan, M. A. U. H., Islam, M. D., Hasan, M. S., Jakir, T., Hossain, A., … & Hasnain, K. N. (2025). Optimizing Solar Energy Production in the USA: Time-Series Analysis Using AI for Smart Energy Management. arXiv preprint arXiv:2506.23368.

[4] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[5] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732.

[6] Bhowmik, P. K., Chowdhury, F. R., Sumsuzzaman, M., Ray, R. K., Khan, M. M., Gomes, C. A. H., … & Gomes, C. A. (2025). AI-Driven Sentiment Analysis for

Bitcoin Market Trends: A Predictive Approach to Crypto Volatility. Journal of Ecohumanism, 4(4), 266–288.

[7] Billah, M., Shatyi, S. S., Sadnan, G. A., Hasnain, K. N., Abed, J., Begum, M., & Sultana, K. S. (2024). Performance Optimization in Multi-Machine Blockchain Systems: A Comprehensive Benchmarking Analysis. Journal of Business and Management Studies, 6(6), 357–375.

[8] Das, B. C., Ahmad, M., & Maqsood, M. (2025). Strategies for Spatial Data Management in Cloud Environments. In Innovations in Optimization and Machine Learning (pp. 181–204). IGI Global Scientific Publishing.

[9] Das, B. C., Mahabub, S., & Hossain, M. R. (2024). Empowering modern business intelligence (BI) tools for data-driven decision-making: Innovations with AI and analytics insights. Edelweiss Applied Science and Technology, 8(6), 8333–8346.

[10] Das, B. C., Zahid, R., Roy, P., & Ahmad, M. (2025). Spatial Data Governance for Healthcare Metaverse. In Digital Technologies for Sustainability and Quality Control (pp. 305–330). IGI Global Scientific Publishing.

[11] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[12] Fariha, N., Khan, M. N. M., Hossain, M. I., Reza, S. A., Bortty, J. C., Sultana, K. S., … & Begum, M. (2025). Advanced fraud detection using machine learning models: enhancing financial transaction security. arXiv preprint arXiv:2506.10842.

[13] Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems, 14(3), 330–347.

[14] Hasan, M. S., Siam, M. A., Ahad, M. A., Hossain, M. N., Ridoy, M. H., Rabbi, M. N. S., … & Jakir, T. (2024). Predictive Analytics for Customer Retention: Machine Learning Models to Analyze and Mitigate Churn in E-Commerce Platforms. Journal of Business and Management Studies, 6(4), 304–320.

[15] Hasanuzzaman, M., Hossain, M., Rahman, M. M., Rabbi, M. M. K., Khan, M. M., Zeeshan, M. A. F., … & Kawsar, M. (2025). Understanding Social Media Behavior in the USA: AI-Driven Insights for Predicting Digital Trends and User Engagement. Journal of Ecohumanism, 4(4), 119–141.

[16] Hossain, M. I., Khan, M. N. M., Fariha, N., Tasnia, R., Sarker, B., Doha, M. Z., … & Siam, M. A. (2025). Assessing Urban-Rural Income Disparities in the USA: A Data-Driven Approach Using Predictive Analytics. Journal of Ecohumanism, 4(4), 300–320.

[17] Islam, M. R., Hossain, M., Alam, M., Khan, M. M., Rabbi, M. M. K., Rabby, M. F., … & Tarafder, M. T. R. (2025). Leveraging Machine Learning for Insights and Predictions in Synthetic E-commerce Data in the USA: A Comprehensive Analysis. Journal of Ecohumanism, 4(2), 2394–2420.

[18] Jakir, T., et al. (2023). Machine Learning-Powered Financial Fraud Detection: Building Robust Predictive Models for Transactional Security. Journal of Economics, Finance and Accounting Studies, 5(5), 161–180.

[19] Khan, M. A. U. H., Islam, M. D., Ahmed, I., Rabbi, M. M. K., Anonna, F. R., Zeeshan, M. D., … & Sadnan, G. M. (2025). Secure Energy Transactions Using Blockchain Leveraging AI for Fraud Detection and Energy Market Stability. arXiv preprint arXiv:2506.19870.

[20] Khan, M. N. M., Fariha, N., Hossain, M. I., Debnath, S., Al Helal, M. A., Basu, U., … & Gurung, N. (2025). Assessing the Impact of ESG Factors on Financial Performance Using an AI-Enabled Predictive Model. International Journal of Environmental Sciences, 1792–1811.

[21] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17), 4768–4777.

[22] Mahabub, S., Das, B. C., & Hossain, M. R. (2024). Advancing healthcare transformation: AI-driven precision medicine and scalable innovations through data analytics. Edelweiss Applied Science and Technology, 8(6), 8322–8332.

[23] Rahman, M. S., Hossain, M. S., Rahman, M. K., Islam, M. R., Sumon, M. F. I., Siam, M. A., & Debnath, P. (2025). Enhancing Supply Chain Transparency with Blockchain: A Data-Driven Analysis of Distributed Ledger Applications. Journal of Business and Management Studies, 7(3), 59–77.

[24] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

[25] Sultana, K. S., Begum, M., Abed, J., Siam, M. A., Sadnan, G. A., Shatyi, S. S., & Billah, M. (2025). Blockchain-Based Green Edge Computing: Optimizing Energy Efficiency with Decentralized AI Frameworks. Journal of Computer Science and Technology Studies, 7(1), 386–408.

[26] United States Census Bureau. (2021). American Community Survey 5-Year Data (2015–2019). Retrieved from https://www.census.gov/programs-surveys/acs.