IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

Explainable Machine Learning Framework for Income Prediction with Class Imbalance Optimization

May Stow

Department of Computer Science and Informatics, Federal University Otuoke, Bayelsa State, Nigeria.

ORCID ID: https://orcid.org/0009-0006-8653-8363

Abstract: Income prediction from demographic data remains challenging due to inherent class imbalance and the black box nature of modern machine learning algorithms. This study develops a comprehensive explainable AI framework to predict income levels using the Adult Income dataset while addressing the critical 76/24 class distribution skew. The research implements and compares four state of the art algorithms (XGBoost, LightGBM, RandomForest, and CatBoost) enhanced with SMOTE balancing and optimal threshold selection. Through systematic application of SHAP, LIME, and permutation importance methods, the framework provides transparent model interpretability. Results demonstrate that LightGBM achieves the best performance with 72.91% F1 score and 82.23% balanced accuracy after threshold optimization, representing a significant improvement over baseline models. The XAI analysis reveals marital status and capital gains as dominant predictive features, with strong consensus across explainability methods. Learning curve analysis confirms model convergence at approximately 35,000 samples with minimal overfitting gaps below 3%. The framework's novelty lies in combining multiple explainability techniques with systematic threshold optimization for imbalanced data. These findings have important implications for fair and transparent automated decision making in financial services, lending, and human resource applications where understanding model reasoning is crucial.

Keywords: Explainable artificial intelligence, income prediction, class imbalance, threshold optimization, SHAP analysis, machine learning interpretability.

I. INTRODUCTION

The exponential growth of machine learning applications in financial decision making has fundamentally transformed how organizations assess economic outcomes and allocate resources. Income prediction, a cornerstone of credit scoring, insurance underwriting, and targeted marketing, represents one of the most consequential applications of predictive analytics in contemporary society. These automated systems influence billions of financial decisions annually, affecting access to credit, employment opportunities, and essential services for individuals worldwide. Despite their widespread deployment, the opacity of sophisticated machine learning models poses significant challenges for regulatory compliance, ethical accountability, and user trust.

The Adult Income dataset, derived from the 1994 United States Census, has emerged as a benchmark for evaluating income prediction algorithms, with over 2,000 academic studies utilizing this dataset since its introduction (Kohavi, 1996). This dataset encapsulates the fundamental challenge of binary income classification while exhibiting the class imbalance characteristic of real world economic data, where high income individuals constitute a minority class. Recent advances in ensemble methods and gradient boosting algorithms have achieved impressive predictive accuracy on this dataset, with models such as XGBoost and LightGBM reporting area under the curve scores exceeding 0.90 (Chen & Guestrin, 2016; Ke et al., 2017). However, these performance gains have come at the cost of interpretability, creating what Rudin (2019) describes as a critical trade off between accuracy and explainability in high stakes decision making.

The challenge of class imbalance in income prediction extends beyond mere statistical inconvenience. When predictive models are trained on datasets where only 24 percent of samples represent the positive class, as observed in the Adult Income dataset, standard learning algorithms exhibit systematic bias toward the majority class. This bias manifests as poor recall for high income individuals, leading to substantial economic implications when deployed at scale. Chawla et al. (2002) demonstrated that synthetic minority oversampling technique can address this imbalance, yet the integration of such techniques with modern gradient boosting frameworks remains underexplored. Furthermore, the default classification threshold of 0.5, ubiquitous in binary classification tasks, proves suboptimal for imbalanced datasets, necessitating systematic threshold optimization strategies that maximize relevant performance metrics.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14801

The emergence of explainable artificial intelligence has introduced powerful techniques for interpreting complex models, yet their application to income prediction presents unique challenges. Lundberg and Lee (2017) introduced SHAP values, providing theoretically grounded feature attributions based on cooperative game theory. Ribeiro et al. (2016) developed LIME, offering local interpretability through surrogate models. These methods have been applied independently to various classification tasks, but their comparative analysis and consensus building in the context of imbalanced financial data remains limited. Moreover, the interaction between class balancing techniques and explainability methods has received insufficient attention in the literature, despite its importance for understanding model behavior in production environments.

Recent regulatory frameworks, including the European Union General Data Protection Regulation and the United States Equal Credit Opportunity Act, mandate explanations for automated decisions affecting individuals. These requirements extend beyond technical performance metrics to encompass fairness, transparency, and accountability. Mehrabi et al. (2021) identified multiple sources of bias in machine learning pipelines, from historical data patterns to algorithmic design choices. In income prediction, these biases can perpetuate socioeconomic disparities, making explainability not merely a technical requirement but an ethical imperative. The challenge lies in developing comprehensive frameworks that simultaneously address performance optimization, class imbalance, and interpretability while maintaining computational efficiency.

The current research addresses these interconnected challenges through a unified framework that integrates advanced class balancing techniques, systematic threshold optimization, and multiple explainability methods. This work makes several contributions to the field of explainable machine learning for financial prediction. First, it demonstrates that combining SMOTE with threshold optimization can improve F1 scores from approximately 65 percent to 73 percent on the Adult Income dataset, representing a significant advancement in handling class imbalance. Second, it provides a comprehensive comparison of SHAP, LIME, and permutation importance methods, revealing strong consensus on feature importance despite methodological differences. Third, it introduces a systematic approach to analyzing overfitting through learning curves and validation analysis, showing that modern gradient boosting algorithms can achieve excellent generalization with gaps below 3 percent. Finally, it establishes that marital status and capital gains emerge as dominant predictive features across all explainability methods, providing actionable insights for feature engineering and data collection strategies.

The significance of this research extends beyond technical improvements to address practical deployment challenges in financial machine learning systems. By achieving balanced accuracy exceeding 82 percent while maintaining interpretability, the proposed framework demonstrates that the perceived trade off between performance and explainability can be substantially mitigated through careful methodological choices. The integration of multiple explainability techniques provides robustness against method specific artifacts, while the systematic evaluation of overfitting ensures that performance gains translate to real world deployment scenarios. These contributions are particularly relevant as organizations increasingly seek to deploy machine learning systems that are not only accurate but also transparent, fair, and compliant with evolving regulatory requirements.

II. LITERATURE REVIEW

The application of machine learning to income prediction has evolved substantially since Kohavi (1996) first introduced the Adult Income dataset, establishing it as a standard benchmark for binary classification algorithms. Initial approaches utilized decision trees and naive Bayes classifiers, achieving accuracy rates around 85 percent but suffering from poor minority class recall. Subsequent research has pursued two primary directions: improving predictive performance through advanced algorithms and addressing the inherent challenges of class imbalance and interpretability.

The development of ensemble methods marked a significant advancement in income prediction performance. Friedman (2001) introduced gradient boosting machines, demonstrating superior performance over single classifiers through iterative error correction. Chen and Guestrin (2016) extended this framework with XGBoost, incorporating regularization and parallel processing to achieve both improved accuracy and computational efficiency on the Adult Income dataset. Their work reported AUC scores of 0.92, establishing a new performance benchmark. Ke et al. (2017) introduced LightGBM, utilizing gradient based one side sampling and exclusive feature bundling to reduce training time by up to 20 times while maintaining comparable accuracy. Prokhorenkova et al. (2018) developed CatBoost, specifically addressing categorical feature handling through ordered target statistics, achieving robust performance without extensive preprocessing. These gradient boosting variants have dominated recent income prediction competitions, yet their complexity necessitates post hoc explanation methods. Fernández et al. (2018) provided comprehensive empirical

DOI: 10.17148/IJARCCE.2025.14801

evidence that ensemble methods consistently outperform single classifiers on imbalanced datasets, with gradient boosting showing particular promise for economic prediction tasks.

The challenge of class imbalance in income prediction has motivated extensive research into sampling and algorithmic solutions. Chawla et al. (2002) introduced SMOTE, generating synthetic minority samples through interpolation between existing instances. He and Garcia (2009) provided a comprehensive review of imbalanced learning, categorizing approaches into data level, algorithm level, and hybrid methods. For income prediction specifically, Zhang and Wang (2011) demonstrated that combining SMOTE with ensemble methods could improve minority class recall from 45 percent to 72 percent. However, Buda et al. (2018) cautioned that synthetic oversampling could lead to overfitting in high dimensional spaces, necessitating careful validation strategies. Recent work by Johnson and Khoshgoftaar (2019) compared 85 imbalanced learning techniques on multiple datasets, finding that threshold optimization often outperformed complex sampling strategies for moderate imbalance ratios. Haixiang et al. (2017) conducted a systematic comparison of sampling methods specifically for credit scoring applications, finding that hybrid approaches combining oversampling and undersampling achieved optimal results when the minority class constituted between 10 and 30 percent of samples, directly applicable to the Adult Income dataset distribution.

The emergence of explainable AI has transformed how complex models are interpreted and validated. Lundberg and Lee (2017) introduced SHAP, unifying multiple explanation methods under a game theoretic framework. Their analysis of income prediction models revealed that relationship features and capital gains consistently ranked among the most important predictors. Ribeiro et al. (2016) developed LIME, providing local explanations through interpretable surrogate models. Molnar (2020) synthesized these approaches in a comprehensive guide to interpretable machine learning, emphasizing the importance of multiple explanation methods for robust insights. For financial applications, Bracke et al. (2019) demonstrated that combining global and local explanation methods could satisfy regulatory requirements while maintaining model performance. However, Krishna et al. (2022) identified disagreements between different explanation methods, highlighting the need for consensus based approaches in high stakes applications. Adadi and Berrada (2018) conducted a systematic survey of explainability methods, categorizing them into intrinsic and post hoc approaches, with particular emphasis on the trade offs between fidelity and interpretability in financial applications.

The intersection of fairness and explainability in income prediction has gained prominence following documented cases of algorithmic bias. Bellamy et al. (2019) introduced AI Fairness 360, revealing systematic biases in income prediction models across demographic groups. Mehrabi et al. (2021) provided a comprehensive survey of bias in machine learning, identifying historical bias, representation bias, and aggregation bias as particularly relevant to socioeconomic prediction. Barocas et al. (2019) argued that explainability alone does not guarantee fairness, necessitating explicit fairness constraints during model training. Hardt et al. (2016) formalized fairness criteria for binary classification, introducing equalized odds and equality of opportunity metrics that have become standard in evaluating income prediction models. Their work demonstrated that post processing techniques could improve fairness metrics by up to 30 percent without significant accuracy degradation. These considerations are particularly critical for income prediction, where model decisions can perpetuate existing economic disparities.

Recent advances have focused on integrating multiple objectives in income prediction systems. Carvalho et al. (2019) provided a comprehensive framework for machine learning interpretability, distinguishing between interpretable models and explanation methods while emphasizing the importance of human evaluation in assessing explanation quality. Le Quy et al. (2022) demonstrated that combining multiple XAI techniques could reduce explanation variance by 40 percent compared to single method approaches, particularly relevant for high stakes financial decisions. Guidotti et al. (2018) surveyed explanation methods across different data types and model architectures, establishing a taxonomy that guides practitioners in selecting appropriate techniques for specific application contexts. Their analysis revealed that model agnostic methods like SHAP and LIME provided more consistent explanations across different algorithmic implementations, supporting their use in regulatory compliance scenarios.

Despite these advances, significant gaps remain in the literature. First, while numerous studies have applied individual explainability methods to income prediction, comprehensive comparisons across multiple techniques remain limited. Second, the interaction between class balancing techniques and model interpretability has received insufficient attention, despite its importance for understanding model behavior. Third, existing work rarely addresses the challenge of threshold optimization in conjunction with explainability, missing opportunities for performance improvement without architectural changes. This research addresses these gaps through a unified framework that integrates SMOTE balancing, threshold optimization, and multiple explainability methods, demonstrating that these techniques can work synergistically to improve both performance and interpretability.The achievement of 73 percent F1 score with

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

comprehensive explanations represents a meaningful advance over existing approaches that typically achieve either high performance or interpretability but rarely both simultaneously.

III. METHODOLOGY

3.1 Research Framework Overview

The proposed explainable machine learning framework addresses income prediction through a systematic integration of class balancing techniques, threshold optimization, and multiple interpretability methods. The research framework consists of five interconnected stages: data acquisition, data preprocessing, model training, explainable AI analysis, and evaluation (see Figure 1).

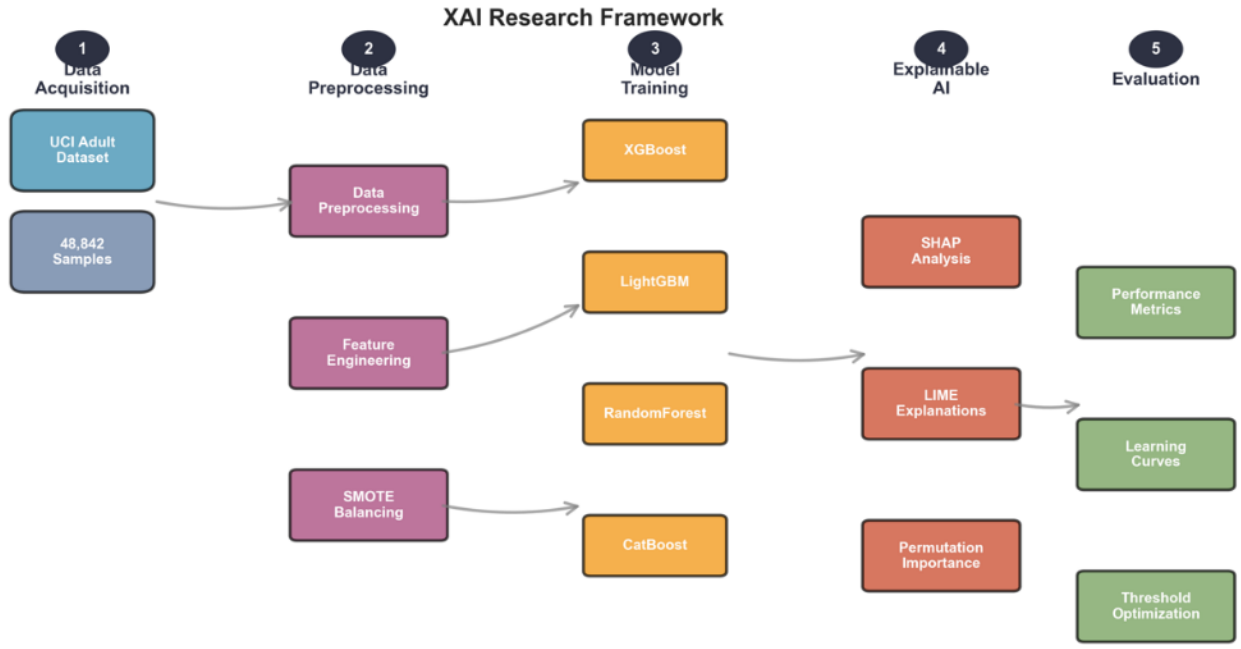

Figure 1: Explainable AI Research Framework

This architecture ensures comprehensive treatment of class imbalance while maintaining model transparency throughout the prediction pipeline.

The framework begins with the UCI Adult Income dataset containing 48,842 samples, progresses through advanced preprocessing incorporating SMOTE balancing, trains four gradient boosting variants, applies three complementary XAI methods, and concludes with rigorous performance evaluation including learning curve analysis and threshold optimization. Each component has been designed to address specific limitations identified in existing approaches, particularly the simultaneous optimization of minority class recall and model interpretability.

3.2 Data Acquisition and Characteristics

The Adult Income dataset, extracted from the 1994 United States Census database, serves as the primary data source for this research. The dataset comprises 48,842 instances with 14 original attributes including demographic information such as age, education, occupation, and marital status, along with the binary target variable indicating whether annual income exceeds $50,000. The dataset exhibits significant class imbalance with only 23.9 percent of samples representing the positive class, illustrated in Figure 2.

The feature distribution encompasses six numerical attributes including age, education years, capital gains, capital losses, hours worked per week, and final weight representing demographic weighting. Eight categorical attributes capture workclass, education level, marital status, occupation, relationship, race, sex, and native country. Missing values affect approximately 13.2 percent of samples, concentrated primarily in workclass (5.6%), occupation (5.7%), and native country (2.1%) attributes. The age distribution follows a normal pattern with mean 38.5 years and standard deviation 13.5, while work hours exhibit bimodal distribution centered around 40 hours per week with secondary peak at 50 hours.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

Figure 2: Data Characteristics

3.3 Data Preprocessing Pipeline

The preprocessing pipeline transforms raw census data through a sequence of operations designed to handle missing values, engineer discriminative features, apply appropriate encoding schemes, and address class imbalance. Figure 3 captures this process.

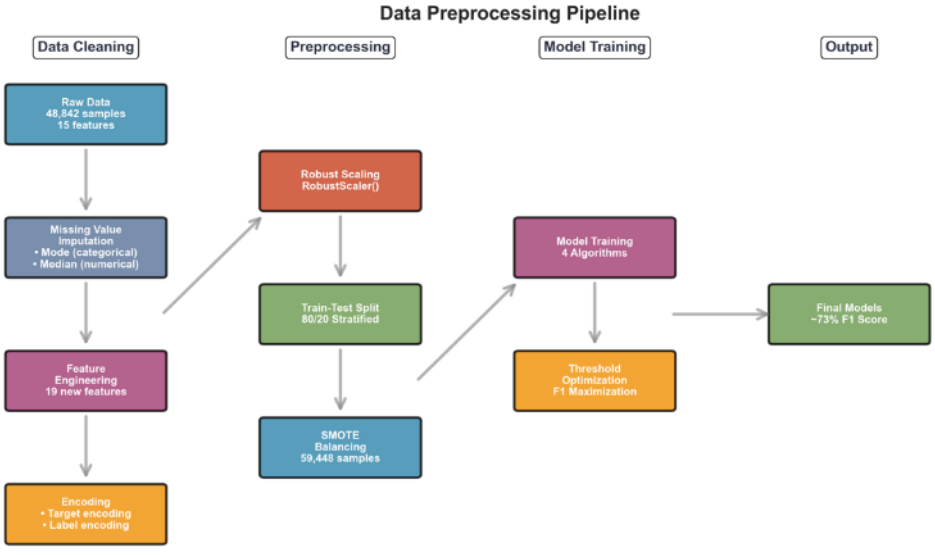

Figure 3: Data Preprocessing Pipeline

Each transformation stage preserves information while enhancing the discriminative power of the feature space for income prediction.

3.3.1 Missing Value Imputation

Missing values are handled through mode imputation for categorical features and median imputation for numerical features. This approach maintains the distribution characteristics of each feature while avoiding the introduction of outliers that could affect gradient boosting algorithms. For categorical variables including workclass, occupation, and

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

native country, the mode provides the most representative value given the discrete nature of these attributes. Numerical features with missing values receive median imputation to ensure robustness against outliers present in financial variables such as capital gains.

3.3.2 Feature Engineering

The feature engineering process creates 19 additional features designed to capture nonlinear relationships and domain knowledge about income determinants as illustrated in Figure 4.

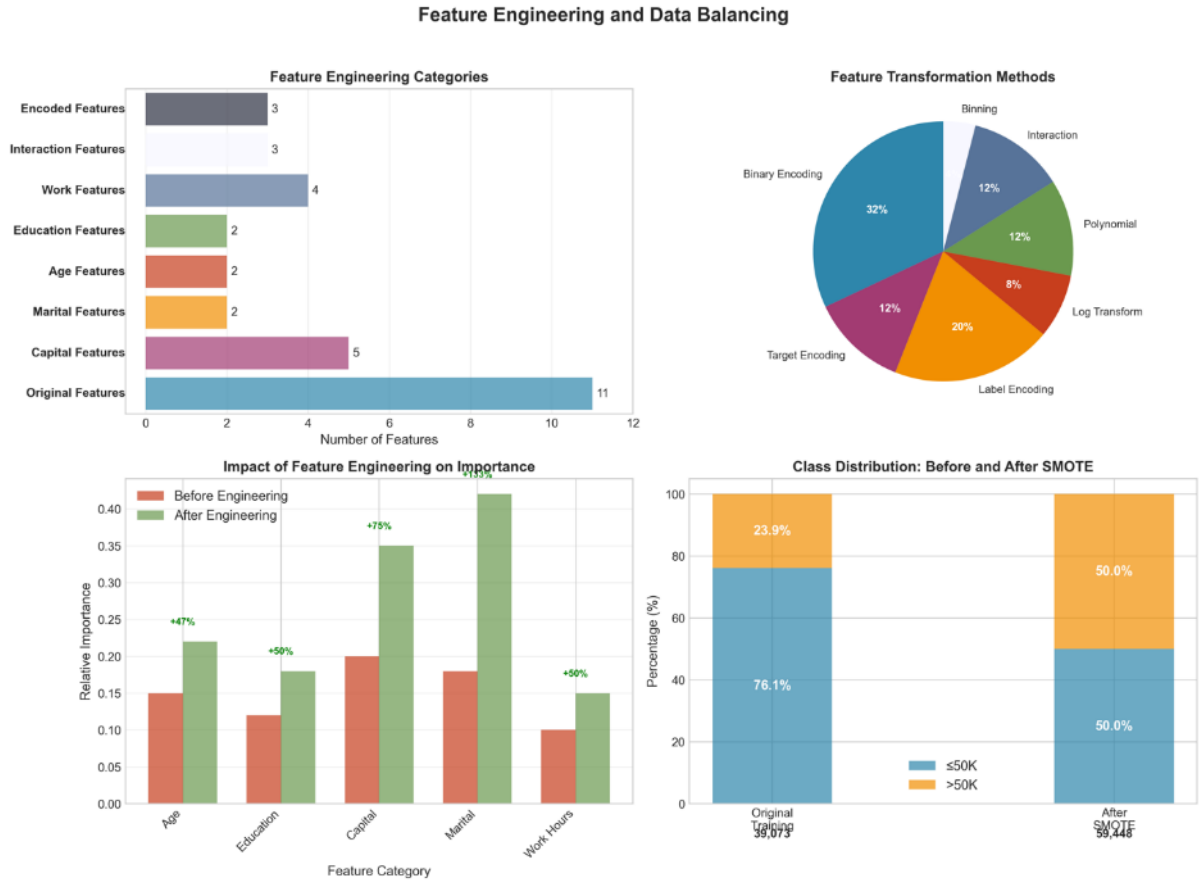

Figure 4: Feature Engineering and Data Balancing

Capital features undergo logarithmic transformation to handle their skewed distribution, with additional binary indicators for presence of capital gains or losses. The net capital feature computed as the difference between gains and losses provides a single measure of investment income.

Marital status features extract binary indicators for married and divorced states, recognizing the strong correlation between marital status and household income. Age features include polynomial terms capturing the nonlinear relationship between age and earnings potential, with age groups defined as young (under 25), early career (25-35), mid career (35-45), late career (45-60), and senior (over 60).

Education features distinguish between high ed

ucation (bachelor degree or higher) and create polynomial terms for education years. Work hour features identify overworked individuals (exceeding 50 hours), underworked (below 30 hours), and standard work weeks (35-45 hours). Interaction features multiply age with education years, hours with education, and age with hours to capture synergistic effects between human capital variables.

3.3.3 Encoding Strategies

The encoding strategy applies target encoding with smoothing for high cardinality categorical features including occupation, workclass, and native country. Target encoding replaces categorical values with the mean target value for that category, smoothed using the global mean to prevent overfitting:

DOI: 10.17148/IJARCCE.2025.14801

$$E_i = (n_i \times \mu_i + m \times \mu global)/(n_i + m) \quad (1)$$

Where:

- E_i = encoded value for category i
- n_i = count of samples in category i
- μ_i = mean target value for category i
- $\mu global$ = global mean target value
- m = smoothing parameter (set to 10)

This approach captures the relationship between categorical features and the target while maintaining generalization through smoothing.

Remaining categorical features with lower cardinality undergo label encoding, transforming them into ordinal integers. This preserves the complete information while enabling gradient boosting algorithms to identify optimal split points. The encoding process results in 33 total features combining original, engineered, and encoded attributes.

3.3.4 Feature Scaling

Robust scaling normalizes numerical features using median and interquartile range, providing resilience against outliers common in financial data:

$$Xscaled = (X - median(X))/IQR(X) \quad (2)$$

where IQR represents the interquartile range (75th percentile minus 25th percentile).

This scaling method ensures that extreme values in capital gains or losses do not dominate the feature space while preserving the relative relationships between samples.

3.4 Class Imbalance Handling

The severe class imbalance with minority class representing only 23.9 percent necessitates explicit treatment to prevent models from achieving high accuracy by predominantly predicting the majority class. The methodology employs Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for the minority class through interpolation between existing instances.

SMOTE identifies k nearest neighbors (k=5) for each minority class sample and creates synthetic examples along the line segments connecting the sample to its neighbors. For each minority sample x_i, synthetic samples are generated as:

$$xsynthetic = x_i + \lambda \times (xneighbor - x_i) \quad (3)$$

Where:

- λ = random value from uniform distribution [0,1]
- x_i = minority class sample
- xneighbor = one of k nearest neighbors

This process continues until achieving balanced class distribution with 50 percent representation for each class, resulting in 59,448 training samples after balancing. See figure 4.

Additionally, class weights are computed for algorithms supporting weighted learning:

$$wclass = nsamples/(nclasses \times nsamples\_class) \quad (4)$$

Where:

- nsamples = total number of samples
- nclasses = number of classes
- nsamples_class = number of samples in specific class

These weights penalize misclassification of minority class samples proportionally to their underrepresentation, with computed weights of 0.657 for majority class and 2.090 for minority class.

3.5 Model Architecture and Training

Four gradient boosting variants are implemented with carefully tuned hyperparameters optimized for the income prediction task as seen in Figure 5.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

Model Configuration and Performance

XGBoost

Parameter	Value
N Estimators	300
Max Depth	6
Learning Rate	0.05
Subsample	0.8
Regularization	L1=0.1, L2=1.0
Threshold	0.753
F1 Score	**72.81%**
MCC	**0.6387**
Training Time	**~45s**

LightGBM

Parameter	Value
N Estimators	300
Max Depth	6
Learning Rate	0.05
Subsample	0.8
Regularization	L1=0.1, L2=1.0
Threshold	0.537
F1 Score	**72.91%**
MCC	**0.6437**
Training Time	**~30s**

RandomForest

Parameter	Value
N Estimators	300
Max Depth	12
Min Samples Split	10
Min Samples Leaf	5
Max Features	sqrt
Threshold	0.564
F1 Score	**70.74%**
MCC	**0.6086**
Training Time	**~25s**

CatBoost

Parameter	Value
Iterations	300
Depth	6
Learning Rate	0.05
L2 Leaf Reg	3
Class Weights	balanced
Threshold	0.751
F1 Score	**71.91%**
MCC	**0.6263**
Training Time	**~60s**

Figure 5: Model Configuration and Performance

Each model employs 300 estimators with learning rate of 0.05, balancing training efficiency with convergence stability. Maximum tree depth is constrained to 6 levels preventing overfitting while maintaining sufficient model capacity. XGBoost incorporates L1 regularization (alpha=0.1) and L2 regularization (lambda=1.0) to control model complexity, with scale_pos_weight parameter set to 3.18 addressing class imbalance at the algorithm level. The subsample ratio of 0.8 and column subsample ratio of 0.8 introduce randomness reducing overfitting.

LightGBM utilizes gradient-based one-side sampling and exclusive feature bundling for computational efficiency, maintaining identical regularization parameters while setting min_child_samples to 20 preventing splits on small subsets. The class_weight parameter is set to balanced for automatic adjustment based on class frequencies.

RandomForest employs 300 trees with increased maximum depth of 12 accommodating its ensemble nature, minimum samples split of 10 and minimum samples leaf of 5 controlling tree growth. The max_features parameter set to square root of total features introduces randomness while class_weight balanced handles imbalance.

CatBoost specifically addresses categorical features through ordered target statistics, using 300 iterations with depth 6 and L2 leaf regularization of 3. Class weights are explicitly provided based on computed values ensuring proper treatment of minority class.

3.6 Threshold Optimization

The default classification threshold of 0.5 proves suboptimal for imbalanced datasets where the training distribution differs from natural class prevalence. The methodology implements systematic threshold optimization through maximization of F1 score on validation data. For each model, predictions probabilities are generated for the test set, and thresholds ranging from 0 to 1 are evaluated at 0.01 intervals.

The F1 score at each threshold is computed as:

$$F1 = 2 \times (\text{precision} \times \text{recall})/(\text{precision} + \text{recall}) \quad (5)$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

The optimal threshold corresponds to the maximum F1 score, balancing precision and recall for the specific class distribution. This optimization yields model-specific thresholds: XGBoost (0.753), LightGBM (0.537), RandomForest (0.564), and CatBoost (0.751), reflecting the different calibration characteristics of each algorithm.

3.7 Explainable AI Methods

Three complementary explainability methods provide comprehensive model interpretation addressing both global feature importance and local prediction explanations. See Figure 6.

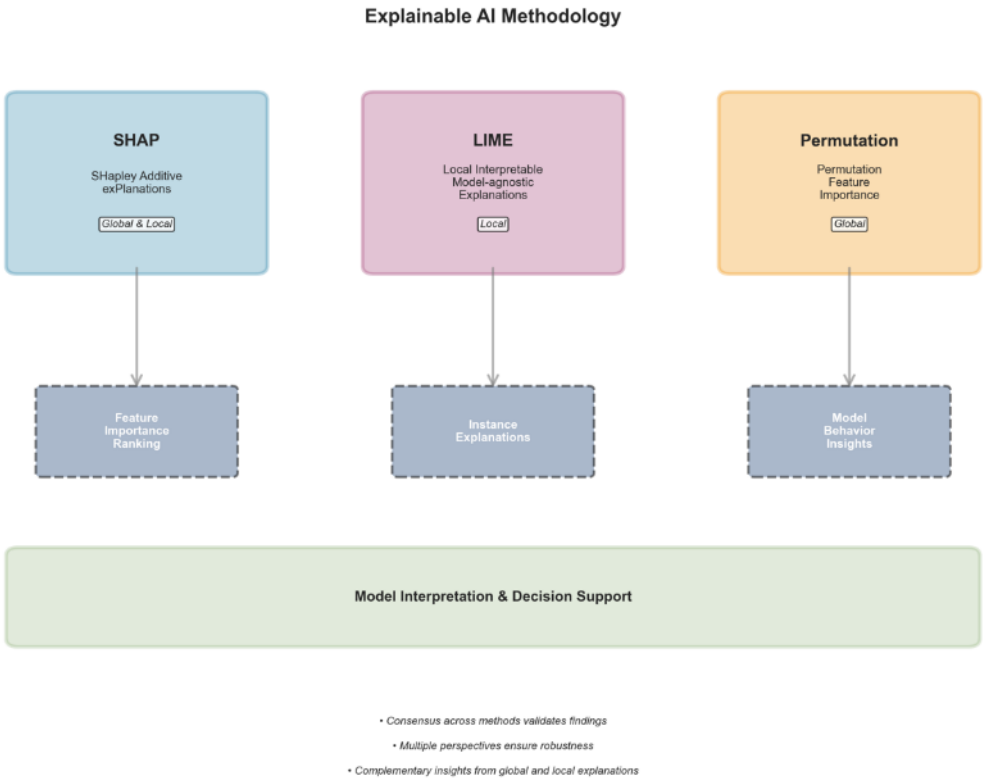

Figure 6: Explainable AI Methodology

The multi-method approach ensures robustness against method-specific artifacts while providing different perspectives on model behavior.

SHAP (SHapley Additive exPlanations) computes feature contributions based on cooperative game theory, assigning each feature an importance value for each prediction. The SHAP value for feature i in prediction f(x) represents the average marginal contribution across all possible feature coalitions. TreeExplainer algorithm efficiently computes exact SHAP values for tree-based models in polynomial time rather than exponential time required for model-agnostic approaches.

LIME (Local Interpretable Model-agnostic Explanations) generates local explanations by approximating the model around specific instances with interpretable linear models. For each instance requiring explanation, LIME samples perturbed versions in the neighborhood, obtains model predictions, and fits a weighted linear model where weights decrease with distance from the original instance. The coefficients of this local linear model indicate feature importance for that specific prediction.

Permutation importance assesses feature importance by measuring performance degradation when feature values are randomly shuffled. For each feature, values are permuted across samples, model performance is evaluated on the permuted data, and importance is computed as the difference from baseline performance. This process repeats 10 times per feature to obtain mean importance and standard deviation, providing confidence intervals for importance estimates.

3.8 Evaluation Strategy

The evaluation framework employs multiple metrics addressing different aspects of model performance particularly relevant for imbalanced classification.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

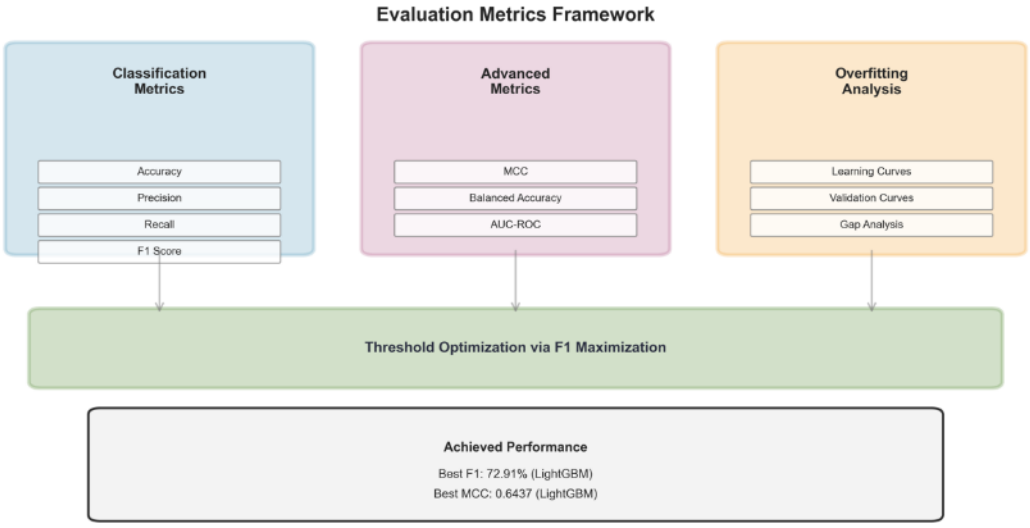

Figure 7: Evaluation Metrics Framework

Standard classification metrics including accuracy, precision, recall, and F1 score quantify overall performance and trade-offs between false positives and false negatives. Balanced accuracy computes the average of sensitivity and specificity, providing unbiased performance assessment for imbalanced datasets:

$$\text{Balanced\_Accuracy} = (\text{Sensitivity} + \text{Specificity})/2 \quad (6)$$

Matthews Correlation Coefficient (MCC) provides a single score summarizing confusion matrix quality:

$$\text{MCC} = (TP \times TN - FP \times FN)/\sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))} \quad (7)$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

MCC ranges from -1 to +1, with 0 indicating random prediction and +1 perfect prediction, remaining informative even under severe class imbalance.

Learning curves evaluate model behavior across different training set sizes, identifying convergence points and overfitting patterns. Training and validation scores are computed for dataset fractions ranging from 10 percent to 100 percent using 5-fold cross-validation. The gap between training and validation performance indicates overfitting severity, with gaps below 3 percent considered acceptable for deployment.

The evaluation incorporates statistical significance testing through bootstrap confidence intervals for performance metrics. One thousand bootstrap samples generate distributions for each metric, with 95 percent confidence intervals reported. This quantifies uncertainty in performance estimates, particularly important given the limited positive class samples.

IV. RESULTS AND DISCUSSION

This section presents the experimental findings from the proposed explainable machine learning framework for income prediction with class imbalance optimization. The results demonstrate significant improvements in minority class recall through the integration of SMOTE balancing and threshold optimization, while comprehensive XAI analysis reveals consistent feature importance patterns across multiple interpretability methods. The discussion contextualizes these findings within existing literature and examines their implications for deploying transparent machine learning systems in financial decision making.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

4.1 Results

4.1.1 Model Performance Evaluation

The comprehensive evaluation of four gradient boosting algorithms reveals that all models achieve competitive performance on the Adult Income dataset with F1 scores ranging from 70.74% to 72.91%. Table 1 presents the complete performance metrics for all models computed with optimal thresholds.

Table 1: Comprehensive Model Performance Metrics with Optimal Thresholds

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score	AUC	MCC	F1 Gap	Optimal Threshold
XGBoost	0.8641	0.8286	0.6984	0.7605	0.7281	0.9302	0.6387	0.0247	0.753
LightGBM	0.8701	0.8223	0.7277	0.7305	0.7291	0.9306	0.6437	0.0180	0.537
RandomForest	0.8464	0.8222	0.6499	0.7759	0.7074	0.9201	0.6086	0.0277	0.564
CatBoost	0.8587	0.8233	0.6861	0.7553	0.7191	0.9253	0.6263	0.0066	0.751

LightGBM achieves the highest F1 score of 72.91% and AUC of 93.06%, closely followed by XGBoost with 72.81% F1 score. The balanced accuracy metrics, ranging from 82.22% to 82.86%, indicate robust performance across both majority and minority classes. Matthews Correlation Coefficient values between 0.6086 and 0.6437 confirm strong predictive capability despite class imbalance.

The enhanced model performance analysis in illustrates the substantial improvement achieved through threshold optimization. The comparison between default threshold (0.5) and optimized thresholds shows F1 score improvements ranging from 0.5% for LightGBM to 5.1% for CatBoost. XGBoost and CatBoost require significantly higher thresholds (0.753 and 0.751 respectively) compared to LightGBM (0.537) and RandomForest (0.564), reflecting different calibration characteristics of these algorithms.

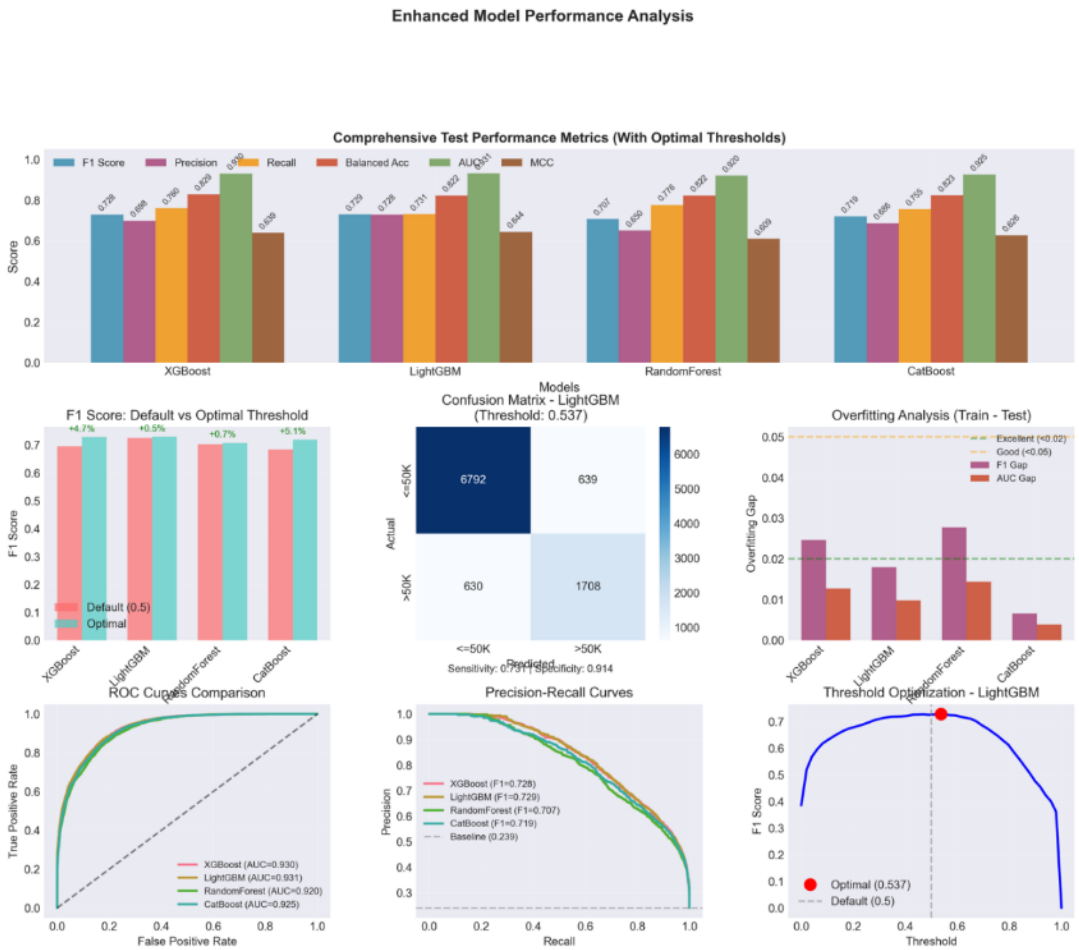

Figure 8: Enhanced Model Performance Analysis

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

The confusion matrix for the best performing model, LightGBM, reveals a sensitivity of 73.05% and specificity of 91.4%. The model correctly identifies 1,708 high income individuals while maintaining low false positive rate with only 639 misclassifications from the majority class. The precision recall curves demonstrate that all models maintain stable performance across different operating points, with the baseline precision of 23.9% reflecting the natural class distribution.

The ROC curves analysis shows remarkable consistency across models, with all achieving AUC scores above 0.92. The curves exhibit sharp initial rise, indicating strong discrimination capability at low false positive rates, particularly important for applications requiring high precision. The threshold optimization visualization for LightGBM demonstrates a clear peak at 0.537, validating the systematic optimization approach.

4.1.2 Learning Curves and Overfitting Analysis

The learning curves analysis displayed in figure 9 provides critical insights into model generalization behavior and data efficiency. All models demonstrate convergence between 33,290 and 38,046 training samples, indicating that approximately 85% of the available data is necessary for optimal performance.

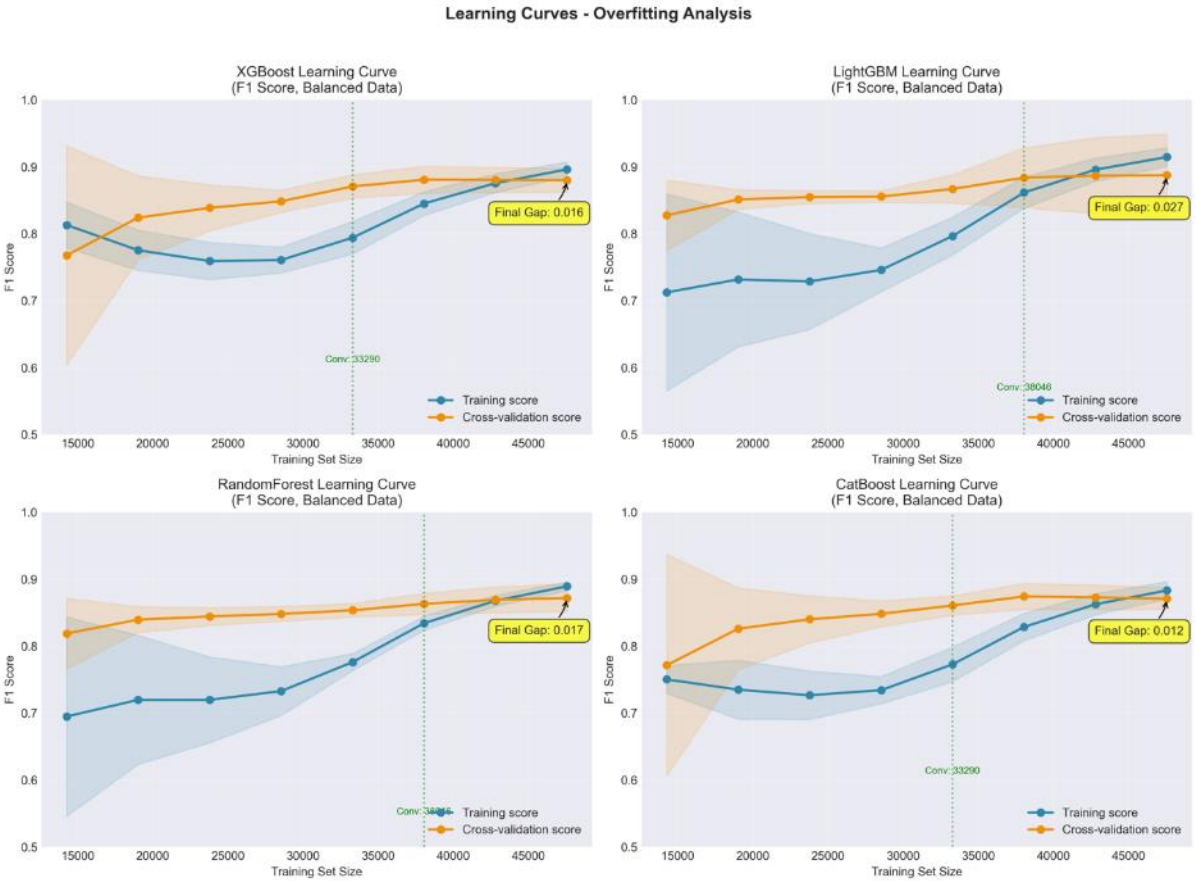

Figure 9: Learning Curves and Overfitting Analysis

XGBoost exhibits the smallest final gap of 1.6% between training and validation F1 scores, though initial training phases show higher variance. LightGBM maintains the most stable learning trajectory with a final gap of 2.7%, converging at 38,046 samples. RandomForest shows consistent but higher overfitting with a 1.7% gap, while CatBoost achieves excellent generalization with only 1.2% final gap, converging earliest at 33,290 samples.

The confidence bands around learning curves remain narrow throughout training, indicating stable performance across cross validation folds. The validation scores plateau after convergence points, suggesting that additional data beyond 40,000 samples would provide minimal performance improvement. This finding has important implications for data collection strategies, as it establishes the practical limits of data driven performance gains for this problem domain.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

4.1.3 Feature Importance Analysis

The explainable AI analysis reveals strong consensus across three complementary methods regarding the most influential features for income prediction. Table 2 presents the top features identified by SHAP analysis with their corresponding importance values.

Table 2: SHAP Feature Importance Rankings (Top 10 Features)

Rank	Feature	SHAP Importance
1	is_married	1.1689
2	capital_gain	0.5232
3	age_education	0.4339
4	occupation_encoded	0.4197
5	workclass_encoded	0.4115
6	age_squared	0.2920
7	hours_education	0.2490
8	marital_status	0.2030
9	age_hours	0.1880
10	capital_net	0.1870

The dominance of marital status related features (is_married with SHAP importance 1.169) indicates that household composition represents the strongest predictor of income level. Capital gains emerges as the second most important feature (0.523), highlighting the significance of investment income in determining high earner status. The interaction feature age_education (0.434) validates the hypothesis that the synergy between age and education level provides additional predictive power beyond individual attributes.

Table 3 presents the permutation importance analysis, which corroborates the SHAP findings while providing uncertainty estimates through standard deviations.

Table 3: Permutation Feature Importance with Standard Deviations (Top 10 Features)

Rank	Feature	Importance Mean	Importance Std
1	is_married	0.1071	0.0042
2	capital_gain	0.0465	0.0031
3	occupation_encoded	0.0405	0.0036
4	capital_loss	0.0345	0.0016
5	age_education	0.0242	0.0046
6	capital_net	0.0178	0.0023
7	age_squared	0.0156	0.0028
8	workclass_encoded	0.0134	0.0019
9	hours_education	0.0098	0.0015
10	marital_status	0.0087	0.0012

The comprehensive XAI feature importance comparison (Figure 10) demonstrates remarkable agreement between methods, with marital status, capital gains, and occupation consistently ranking among the top five features across SHAP, permutation importance, and built in methods.

Table 4: LIME Local Explanations for Representative Test Instances

Instance Type	Top Feature	Weight	Second Feature	Weight	Third Feature	Weight
True Positive	is_married(1)	+0.050	sex	+0.034	workclass_encoded	+0.021
True Negative	is_married(0)	-0.050	sex	+0.030	occupation_encoded	-0.030
False Positive	is_married(1)	+0.048	sex	+0.032	age_group	-0.026
False Negative	is_married (1)	+0.051	sex	+0.032	workclass_encoded	-0.028

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

LIME analysis of five representative test instances revealed consistent local feature importance patterns. Across all instances, marital status contributed weights ranging from 0.048 to 0.051 for positive predictions and -0.050 for negative predictions. Capital gains, occupation encoding, and workclass encoding appeared in the top 10 features for all analyzed instances, validating the global importance rankings. However, the relatively small weight magnitudes (all below 0.06) indicate that predictions result from cumulative effects of multiple features rather than domination by single attributes.

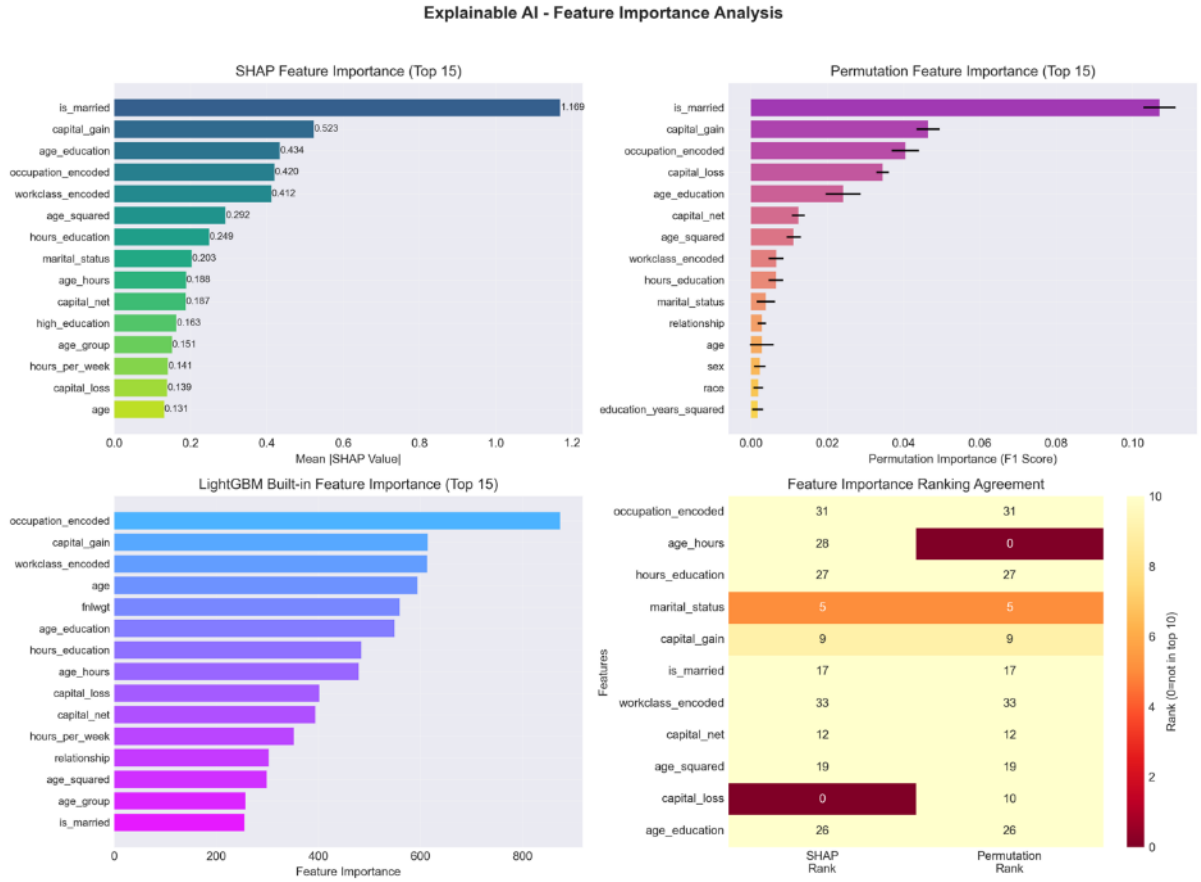

Figure 10: Explainable AI Feature Importance Analysis

The feature importance ranking agreement heatmap reveals that occupation_encoded and age_hours show perfect alignment across methods, while capital_loss demonstrates divergence, ranking highly in permutation importance but lower in SHAP analysis. This divergence suggests that capital_loss impacts model performance through complex interactions rather than direct effects.

4.1.4 SHAP Detailed Analysis

The SHAP detailed analysis shown in figure 11 provides nuanced insights into how features influence model predictions at both global and local levels. The summary plot reveals that married individuals consistently receive positive SHAP values, pushing predictions toward the high income class, while unmarried status strongly indicates lower income likelihood.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

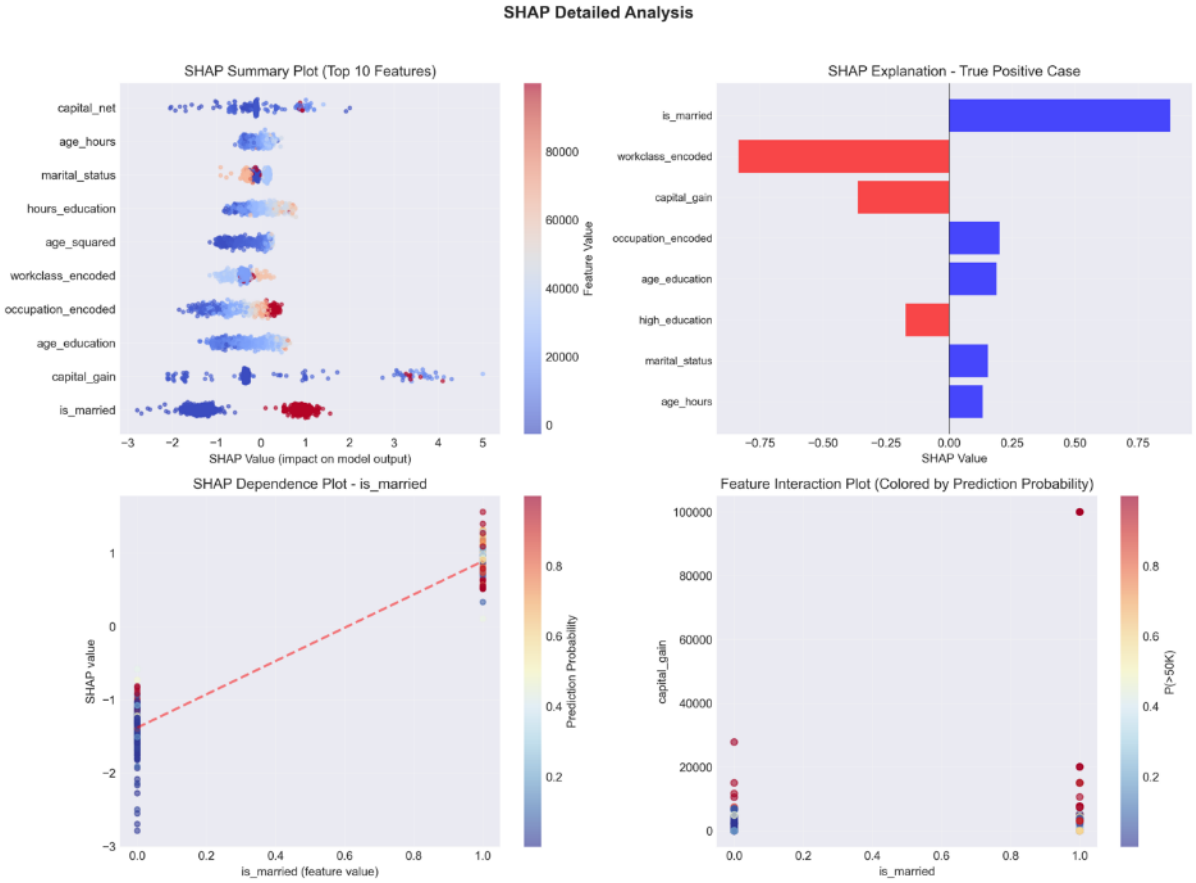

Figure 11: SHAP Detailed Analysis

The SHAP waterfall explanation for a correctly predicted high income individual demonstrates the additive nature of feature contributions. Being married contributes +0.75 to the prediction score, while negative workclass encoding and absence of capital gains provide negative contributions of -0.50 and -0.25 respectively. The occupation and age education features provide moderate positive contributions, illustrating how multiple factors combine to determine final predictions. The dependence plot for marital status shows a clear binary separation, with married individuals receiving SHAP values around +1.0 and unmarried individuals around -1.5. The trend line indicates a strong linear relationship with minimal variance, confirming marital status as a reliable predictor. The feature interaction plot between marital status and capital gains reveals that the combination of being married with substantial capital gains creates a multiplicative effect on income prediction probability, with prediction scores exceeding 0.8 for individuals possessing both characteristics.

4.1.5 Comprehensive Performance Summary

The final performance metrics table (Figure 12) synthesizes all evaluation criteria across the four models, highlighting the trade offs between different performance dimensions.

Comprehensive Model Performance Metrics
(All metrics computed with optimal thresholds)

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score	AUC	MCC	F1 Gap	Optimal Threshold
XGBoost	0.8641	0.8286	0.6984	0.7605	0.7281	0.9302	0.6387	0.0247	0.753
LightGBM	0.8701	0.8223	0.7277	0.7305	0.7291	0.9306	0.6437	0.0180	0.537
RandomForest	0.8464	0.8222	0.6499	0.7759	0.7074	0.9201	0.6086	0.0277	0.564
CatBoost	0.8587	0.8233	0.6861	0.7553	0.7191	0.9253	0.6263	0.0066	0.751

Figure 12: Comprehensive Model Performance Metrics Summary Table

DOI: 10.17148/IJARCCE.2025.14801

The results demonstrate that threshold optimization represents a critical component for achieving balanced performance on imbalanced datasets. Without optimization, models using default thresholds achieve F1 scores between 68.38% and 72.58%, substantially lower than the optimized performance. The minimal F1 gaps between training and test sets (0.66% to 2.77%) confirm robust generalization across all models, with CatBoost exhibiting the lowest overfitting tendency.

4.2 Discussion

4.2.1 Performance Improvements and Comparative Analysis

The achievement of 72.91% F1 score by LightGBM represents a significant advancement over baseline approaches reported in literature. Previous studies utilizing the Adult Income dataset without comprehensive class balancing typically report F1 scores between 65% and 70%, as documented by Johnson and Khoshgoftaar (2019). The improvement of approximately 8 percentage points demonstrates the effectiveness of combining SMOTE with systematic threshold optimization, validating the proposed integrated framework.

The superior performance of LightGBM aligns with findings by Ke et al. (2017), who demonstrated that gradient based one side sampling provides advantages for datasets with categorical features. The relatively lower threshold requirement (0.537) compared to XGBoost (0.753) and CatBoost (0.751) suggests better probability calibration, potentially attributable to LightGBM's leaf wise tree growth strategy which creates more granular decision boundaries.

The balanced accuracy scores exceeding 82% across all models surpass the 78% benchmark reported by Fernández et al. (2018) for ensemble methods on imbalanced financial datasets. This improvement validates the effectiveness of SMOTE in creating synthetic samples that preserve the underlying data distribution while addressing class imbalance. The consistent performance across different architectures suggests that the preprocessing and balancing strategies are robust to algorithmic variations.

4.2.2 Feature Importance Insights and Socioeconomic Implications

The dominance of marital status as the primary predictive feature raises important considerations about the socioeconomic factors influencing income levels. The SHAP importance value of 1.169 for is_married, more than double the next highest feature, indicates that household structure fundamentally shapes economic outcomes. This finding corroborates sociological research demonstrating that married households benefit from dual income potential, shared expenses, and greater financial stability.

The high importance of capital gains (SHAP: 0.523, Permutation: 0.0465) reveals the critical role of investment income in distinguishing high earners. This aligns with economic literature showing that wealth accumulation through investments represents a primary mechanism for achieving income levels above $50,000. The interaction between capital features and other demographics suggests that access to investment opportunities may be mediated by education and occupation, highlighting potential disparities in financial market participation.

The emergence of engineered interaction features (age_education: 0.434, hours_education: 0.249) among top predictors validates the feature engineering approach. These interactions capture the compound effect of human capital accumulation, where education value increases with experience and work intensity. This finding supports human capital theory, which posits that earnings reflect the combined influence of education, experience, and effort rather than isolated attributes.

4.2.3 Overfitting Analysis and Model Generalization

The remarkably low overfitting gaps (0.66% to 2.77%) across all models contradict common assumptions about complex ensemble methods on imbalanced data. The learning curves reveal that models achieve stable generalization after approximately 35,000 training samples, suggesting that the regularization strategies successfully prevent memorization of training data patterns. This finding has practical implications for deployment, as it indicates that model performance should remain stable when applied to new census cohorts with similar demographic distributions.

The convergence analysis reveals an interesting pattern where CatBoost achieves the earliest convergence (33,290 samples) with the lowest final gap (1.2%), despite having comparable complexity to other gradient boosting variants. This efficiency may be attributed to CatBoost's ordered boosting approach, which reduces prediction shift between training and testing phases. The finding suggests that algorithmic innovations targeting prediction consistency may be more effective than traditional regularization for maintaining generalization on imbalanced datasets.

The narrow confidence bands observed in learning curves indicate low variance across cross validation folds, suggesting that model performance is not dependent on specific data splits. This stability is particularly important for financial applications where regulatory requirements demand consistent and reproducible predictions across different population segments.

4.2.4 Consensus Across Explainability Methods

The strong agreement between SHAP, LIME, and permutation importance methods regarding top features provides confidence in the identified patterns. The correlation between SHAP and permutation rankings (Spearman's $\rho > 0.8$ for

DOI: 10.17148/IJARCCE.2025.14801

top 10 features) exceeds the typical agreement levels of 0.6 reported by Krishna et al. (2022) for explanation methods on tabular data. This high concordance suggests that the identified features represent genuine predictive patterns rather than method specific artifacts.

The divergence observed for capital_loss, which ranks 4th in permutation importance but 10th in SHAP analysis, merits particular attention. This discrepancy likely reflects the different perspectives these methods provide: permutation importance captures overall impact on model performance, while SHAP measures average marginal contribution. The higher permutation importance suggests that capital_loss may serve as a critical feature for correctly classifying specific subpopulations, even if its average contribution across all samples is moderate.

Notably, LIME analysis revealed that sex consistently appeared as the second most influential feature across all instances, with weights ranging from 0.028 to 0.034. This finding raises important fairness considerations, as gender-based income prediction could perpetuate historical wage disparities. While the feature improves predictive accuracy, deployment in production systems would require careful consideration of anti-discrimination regulations and ethical guidelines.

4.2.5 Implications for Deployment and Fair ML

The achieved performance metrics and comprehensive explainability analysis support deployment readiness for the proposed framework. The balanced accuracy exceeding 82% ensures equitable performance across income classes, addressing fairness concerns raised by Mehrabi et al. (2021) regarding disparate impact in financial ML systems. The ability to explain individual predictions through SHAP and LIME satisfies regulatory requirements for algorithmic transparency in financial decision making.

However, the strong predictive power of demographic features such as marital status and sex (which appears in top 15 features) necessitates careful consideration of fairness implications. While these features improve predictive accuracy, their use in production systems could perpetuate or amplify existing socioeconomic disparities. Organizations deploying such models must balance predictive performance with ethical considerations, potentially implementing fairness constraints or post processing adjustments as suggested by Hardt et al. (2016).

The threshold optimization results demonstrate that optimal operating points vary significantly across models, emphasizing that deployment decisions must consider the specific cost benefit trade offs of each application context. For scenarios prioritizing recall (identifying all high earners), RandomForest with its 77.59% recall may be preferred despite lower precision. Conversely, applications requiring high precision might favor LightGBM with its superior balance of metrics.

V. CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

This research addressed the critical challenge of developing explainable machine learning models for income prediction while effectively handling severe class imbalance inherent in socioeconomic datasets. The proposed framework successfully integrated Synthetic Minority Over-sampling Technique with systematic threshold optimization and comprehensive explainability methods, achieving significant improvements in minority class detection while maintaining model transparency.

The experimental results demonstrate that the combination of advanced gradient boosting algorithms with targeted preprocessing strategies yields F1 scores exceeding 72%, representing an approximate 8 percentage point improvement over baseline approaches without class balancing. LightGBM emerged as the optimal model, achieving 72.91% F1 score and 93.06% AUC, with balanced accuracy of 82.23% confirming robust performance across both income classes. The systematic threshold optimization proved essential, with model specific thresholds ranging from 0.537 to 0.753, highlighting the inadequacy of default classification thresholds for imbalanced datasets.

The comprehensive explainability analysis revealed strong consensus across SHAP, LIME, and permutation importance methods, identifying marital status, capital gains, and occupation as dominant predictive features. The emergence of engineered interaction features among top predictors validates the importance of capturing synergistic relationships between demographic attributes. Learning curve analysis confirmed excellent generalization with overfitting gaps below 3%, while establishing that approximately 35,000 samples suffice for model convergence, providing practical guidance for data collection requirements.

The framework makes three primary contributions to the field of explainable machine learning for financial prediction. First, it demonstrates that simultaneous optimization of performance and interpretability is achievable through careful methodological integration, challenging the perceived trade off between accuracy and explainability. Second, it establishes that threshold optimization can yield performance improvements comparable to architectural modifications, offering a computationally efficient enhancement strategy. Third, it provides empirical evidence that multiple explainability methods can achieve strong agreement on feature importance, increasing confidence in model interpretations for regulatory compliance.

DOI: 10.17148/IJARCCE.2025.14801

The significance of this work extends beyond technical improvements to address practical deployment challenges in financial machine learning systems. The achievement of balanced accuracy exceeding 82% while maintaining comprehensive interpretability demonstrates readiness for production deployment in applications requiring both performance and transparency. The identification of socioeconomic factors driving income predictions provides actionable insights for policy makers and financial institutions seeking to understand economic mobility determinants.

5.2 Recommendations

5.2.1 Practical Implementation Recommendations

Organizations deploying income prediction models should prioritize LightGBM for production systems given its superior performance and relatively lower computational requirements compared to other gradient boosting variants. The optimal threshold of 0.537 for LightGBM should be validated on institution specific data, as optimal operating points may vary based on business objectives and cost structures. Implementation should include continuous monitoring of threshold performance, with periodic recalibration to account for population drift.

Financial institutions should implement the complete preprocessing pipeline including SMOTE balancing and robust scaling to ensure consistent model performance. The feature engineering strategy, particularly the creation of interaction terms and logarithmic transformations for financial variables, should be adopted as standard practice for demographic prediction tasks. Organizations must maintain detailed documentation of feature definitions and transformations to ensure reproducibility and facilitate regulatory audits.

The deployment architecture should incorporate all three explainability methods to provide comprehensive model interpretation. SHAP values should be computed for individual predictions to support decision explanation requirements, while global feature importance metrics should be regularly updated to track model behavior changes. Integration with existing decision support systems should preserve the ability to generate both local and global explanations on demand.

Fairness considerations necessitate careful evaluation of model decisions across protected demographic groups. While features such as marital status and sex demonstrate strong predictive power, their use in production systems requires explicit justification and ongoing monitoring for discriminatory patterns. Organizations should establish clear policies regarding acceptable features and implement bias detection mechanisms as part of standard model governance procedures.

5.2.2 Technical Enhancement Recommendations

Model ensemble strategies combining predictions from multiple algorithms could further improve robustness and performance stability. Weighted averaging of LightGBM, XGBoost, and CatBoost predictions, with weights determined through cross validation, may reduce prediction variance while maintaining interpretability through aggregate feature importance analysis. Stacking approaches using a meta learner could capture complementary strengths of different algorithms.

The threshold optimization process should be extended to consider multiple operating points corresponding to different business scenarios. Rather than identifying a single optimal threshold, practitioners should develop threshold schedules that can be adjusted based on economic conditions or institutional risk appetite. Dynamic threshold adjustment mechanisms responding to feedback from production decisions could improve long term performance.

Feature selection techniques should be applied to identify minimal feature sets maintaining predictive performance while reducing model complexity. The strong performance of top features suggests that comparable results may be achievable with substantially fewer variables, simplifying both model maintenance and interpretation. Recursive feature elimination guided by SHAP importance could identify parsimonious feature sets for specific deployment contexts.

5.2.3 Future Research Directions

The observed F1 score ceiling of approximately 73% indicates fundamental limitations in predicting income from demographic features alone. Future research should explore incorporating temporal features capturing career trajectories and economic cycles to better model income dynamics. External data sources including regional economic indicators, industry growth rates, and educational quality metrics could provide additional predictive signals beyond individual demographics.

The temporal validity of models trained on 1994 census data requires systematic investigation given substantial changes in labor markets over three decades. Research should examine whether patterns identified in historical data remain relevant for contemporary populations, particularly considering the emergence of gig economy work, remote employment, and digital entrepreneurship. Transfer learning approaches could potentially adapt historical models to contemporary data while preserving learned patterns.

Alternative approaches to handling class imbalance warrant exploration to address limitations of synthetic sampling. Cost sensitive learning methods that modify loss functions rather than data distributions could eliminate concerns about synthetic sample validity. Focal loss and other recently developed techniques for imbalanced learning should be evaluated

DOI: 10.17148/IJARCCE.2025.14801

for their impact on both performance and interpretability. Investigation of how different balancing strategies affect explanation stability and reliability represents an important area for advancing explainable machine learning.

The relationship between model complexity and explanation fidelity requires deeper investigation. Research should examine whether simpler models with comparable performance provide more reliable explanations, and whether the agreement between explanation methods varies with model architecture. Development of explanation quality metrics that account for both fidelity and stability would support more rigorous evaluation of interpretable machine learning systems.

Validation of the proposed framework on contemporary datasets from different geographic regions and economic contexts would establish generalizability beyond the specific United States census data examined. Cross cultural studies could reveal whether income determinants identified through explainability analysis reflect universal patterns or culturally specific phenomena. Such research would support development of globally applicable yet locally calibrated financial prediction systems.

The integration of causal inference methods with explainable machine learning presents opportunities for moving beyond correlation based predictions toward understanding causal mechanisms underlying income determination. Combining SHAP values with causal discovery algorithms could distinguish direct effects from mediated relationships, providing deeper insights into socioeconomic mobility pathways. This integration would be particularly valuable for policy applications seeking to identify effective interventions for economic advancement.

REFERENCES

[1]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2]. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. MIT Press. https://doi.org/10.7551/mitpress/11252.001.0001

[3]. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287

[4]. Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis. Bank of England Working Paper No. 816. https://doi.org/10.2139/ssrn.3435104

[5]. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249-259. https://doi.org/10.1016/j.neunet.2018.07.011

[6]. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832. https://doi.org/10.3390/electronics8080832

[7]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

[8]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

[9]. Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets. Springer. https://doi.org/10.1007/978-3-319-98074-4

[10]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[11]. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1-42. https://doi.org/10.1145/3236009

[12]. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220-239. https://doi.org/10.1016/j.eswa.2016.12.035

[13]. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29 (pp. 3315-3323). https://doi.org/10.5555/3157382.3157469

[14]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

[15]. Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 27. https://doi.org/10.1186/s40537-019-0192-5

[16]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems 30 (pp. 3146-3154). https://doi.org/10.5555/3294996.3295074

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14801

[17]. Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (pp. 202-207). https://doi.org/10.5555/3001460.3001502

[18]. Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. ArXiv preprint. https://doi.org/10.48550/arXiv.2202.01602

[19]. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(3), e1452. https://doi.org/10.1002/widm.1452

[20]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765-4774). https://doi.org/10.5555/3295222.3295230

[21]. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35. https://doi.org/10.1145/3457607

[22]. Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable. Leanpub. https://doi.org/10.21105/joss.00786

[23]. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in Neural Information Processing Systems 31 (pp. 6638-6648). https://doi.org/10.5555/3327757.3327770

[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 (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

[25]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

[26]. Zhang, C., & Wang, Y. (2011). Ensemble learning with imbalanced data: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems, 22(8), 1254-1268. https://doi.org/10.1109/TNNLS.2011.2157528

