International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 10 Issue 5, Sep-Oct 2026 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Study of Explainable Machine Learning Model for Trustworthy Decision Support in GST Revenue Prediction

Prashant Chandra[1], Adarsh Raushan[2], Dr. Vivek Richhariya[3]

[1]M Tech Scholar, [2]Assistant Professor, [3]HOD & Professor,
[1,2,3]CSE Department, LNCT, Bhopal, Madhya Pradesh, India

ABSTRACT

GST is an important source of public revenue in India. Accurate forecasting of GST is important for budgeting, fiscal risk assessment, expenditure planning and policy formulation. In this research we provide an explainable machine learning framework for dependable GST revenue forecasting by integrating SHapley Additive exPlanations (SHAP) and Extreme Gradient Boosting (XGBoost). Historical GST collections were combined with economic, fiscal, administrative, policy and temporal elements. The process included data cleansing, missing-value treatment, outlier analysis, feature engineering, chronological division for training, validation and testing, hyperparameter optimisation, cross-validation, comparison modelling and explainability assessment. The suggested model was compared with Linear Regression, Decision Tree, Support Vector Regression, Random Forest and Gradient Boosting utilising Mean Absolute Error, Root Mean Squared Error, Mean Absolute Percentage Error and coefficient of determination. The results of the trials indicated that the best performing model overall was the XGBoost–SHAP model with an MAE of ₹141.82 crore, RMSE of ₹196.47 crore, MAPE of 4.36% and R2 of 0.928. The cross-validation R2 of 0.918 and the little performance difference between training and testing indicated stable generalisation with controlled overfitting. The SHAP analysis found that the greatest indicators were GST income from the previous month, taxable turnover, return-filing rate, GDP growth and registered taxpayers. The explanations made explicit how each variable had a role in each prediction and in what direction. The proposed framework is a suitable decision-support tool for revenue monitoring, budget preparation, compliance planning and fiscal policy evaluation due to its forecast accuracy, transparency, interpretability and stability. It is designed to supplement, not to replace, professional advice. It is sensitive to change in economic, budgetary and administrative situations.

How to cite this paper: Prashant Chandra | Adarsh Raushan | Dr. Vivek Richhariya "Study of Explainable Machine Learning Model for Trustworthy Decision Support in GST Revenue Prediction" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-10 | Issue-5, October 2026, pp.413-419, URL: www.ijtsrd.com/papers/ijtsrd142131.pdf

IJTSRD142131

Copyright © 2026 by author (s) and International Journal of Trend in Scientific Research and Development Journal. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) (http://creativecommons.org/licenses/by/4.0)

KEYWORDS: *GST Revenue Prediction; Explainable Artificial Intelligence; XGBoost; SHAP; Trustworthy Decision Support; Fiscal Forecasting.*

I. INTRODUCTION

Goods and Services Tax (GST) has become a major source of government revenue in India, making accurate revenue forecasting essential for fiscal planning, budget preparation, resource allocation, and policy formulation. However, GST revenue is influenced by several dynamic factors, including historical collections, economic conditions, taxpayer compliance, business activity, seasonal variations, and policy changes. The complex and nonlinear relationships among these factors create challenges for traditional forecasting approaches. Machine learning (ML) techniques provide improved capability for modelling complex patterns; however, their black-box nature limits transparency and trust in high-stakes financial decision-making. This study proposes an explainable machine learning framework based on Extreme Gradient Boosting (XGBoost) integrated with SHapley Additive exPlanations (SHAP) for trustworthy GST revenue prediction.

The major contributions of this paper are:

A. Development of an XGBoost–SHAP-based forecasting framework that combines prediction accuracy with interpretability.

B. Integration of economic, administrative, and temporal variables for improved GST revenue modelling.

C. Comparative evaluation of the proposed model with conventional machine learning approaches using standard performance metrics.

D. Identification of key revenue-driving factors through SHAP-based explainability analysis to support transparent decision-making.

The remainder of the paper is organized as follows. Section I introduces the research background, motivation, and objectives. Section II presents the literature review and identifies existing research gaps. Section III describes the proposed methodology, including data preprocessing, model development, and explainability techniques. Section IV discusses the experimental setup and implementation details. Section V presents the results, comparative analysis, and SHAP interpretation. Finally, Section VI concludes the paper and highlights future research directions for advanced AI-based fiscal forecasting systems.

II. LITERATURE SURVEY

The advancement of artificial intelligence (AI) and machine learning (ML) has significantly transformed financial and economic forecasting by enabling data-driven analysis of complex and nonlinear relationships. Traditional statistical forecasting techniques, such as regression models and autoregressive approaches, have been widely used for revenue prediction; however, these methods often depend on assumptions of linearity and may fail to capture dynamic interactions among multiple economic and administrative factors. Recent studies have therefore focused on machine learning-based forecasting techniques capable of extracting hidden patterns from large-scale datasets and improving prediction accuracy.

Among various ML algorithms, ensemble learning approaches have demonstrated strong performance in financial forecasting applications. Extreme Gradient Boosting (XGBoost), introduced by Chen and Guestrin, has gained considerable attention because of its high prediction capability, computational efficiency, regularization mechanism, and ability to model complex nonlinear relationships in structured datasets [1]. XGBoost-based models have been successfully applied in areas such as financial risk prediction, economic forecasting, fraud detection, and decision-support systems, where they have shown improved performance compared with conventional statistical approaches [2].

However, the increasing complexity of machine learning models has created challenges regarding transparency and interpretability. Many advanced ML models provide highly accurate predictions but function as black-box systems, making it difficult for users to understand the reasons behind specific outputs [3]. In applications such as public finance, taxation, and government policy planning, prediction accuracy alone is insufficient; decision-makers require reliable explanations, accountability, and confidence in AI-generated outcomes. Explainable Artificial Intelligence (XAI) has emerged as an important research area to address these challenges by improving the interpretability and trustworthiness of machine learning models [4].

Among various XAI techniques, SHapley Additive exPlanations (SHAP) has become one of the most widely adopted approaches due to its strong theoretical foundation based on cooperative game theory. Lundberg and Lee introduced SHAP as a unified framework for interpreting machine learning predictions by assigning contribution values to individual input features [5]. SHAP provides both global and local explanations, allowing researchers to identify the most influential variables and understand how each feature contributes positively or negatively to a particular prediction [6]. Recent studies have demonstrated the effectiveness of SHAP-based approaches in financial forecasting, credit risk assessment, fraud detection, and corporate performance prediction [7].

Recent research has also highlighted the importance of trustworthy AI-based decision-support systems in financial and governmental applications [8]. Explainable machine learning models help bridge the gap between predictive performance and human understanding by allowing domain experts to validate model behaviour and integrate AI outputs with professional judgement. Studies have emphasized that transparency, robustness, fairness, and interpretability are essential requirements for deploying AI systems in high-impact decision-making environments [9].

In financial time-series forecasting, recent explainable ML frameworks have demonstrated that combining high-performance algorithms with interpretability methods can improve stakeholder confidence and practical adoption [10]. XGBoost–SHAP-based models have been particularly effective because XGBoost provides strong predictive performance, while SHAP explains the underlying decision process. Such hybrid frameworks have been applied successfully for financial distress prediction, stock market analysis, and economic risk assessment [11].

Although significant progress has been achieved in explainable financial forecasting, limited research has investigated explainable machine learning approaches for GST revenue prediction. Existing GST forecasting studies mainly focus on improving prediction accuracy using statistical and machine learning methods, while limited attention has been given to transparency, feature contribution analysis, and trustworthy decision support. GST revenue is influenced by multiple interacting factors, including previous revenue trends, taxable turnover, economic growth, taxpayer compliance, return filing behavior, consumption patterns, and policy changes. Therefore, an explainable forecasting framework is required to support accurate and transparent fiscal decision-making.

This study addresses this research gap by proposing an XGBoost–SHAP-based framework for GST revenue prediction. The proposed model integrates machine learning-based forecasting with SHAP-driven interpretation to identify key revenue determinants and provide meaningful explanations for prediction outcomes. The framework aims to achieve a balance between forecasting accuracy, interpretability, and reliability, thereby supporting trustworthy AI applications in public finance and revenue management.

III. PROBLEM IDENTIFICATION

Governments use the GST revenue for public spending, budgeting, fiscal planning and policy formation. However, it is difficult to predict this since collections are impacted by administrative, seasonal, economic and policy variables [3]. But these sophisticated machine learning approaches are typically hidden and ordinary forecasting methods generally don't explain these complicated non-linear relationships adequately [4]. Therefore, this study seeks to satisfy the gap in the requirement for an accurate and understandable way of computing GST revenue.

- GST revenues are affected by previous revenue, inflation, imports, consumption, firm activity, compliance, seasonal demand and policy changes [2].
- Traditional approaches are not able to detect hidden, nonlinear and dynamic links in GST revenue data [5].
- The appropriate machine learning algorithms are necessary to increase the accuracy of forecasts and to predict future revenue growth, decrease or shortcomings [12].

Hence, beyond than giving a numerical estimate to policy makers, there is a need for model explainability to explain why the model forecasts an increase or decrease in GST income. The project will create and validate an XGBoost–SHAP model to give accurate, clear and dependable support in tax policy, compliance management and fiscal planning.

IV. RESEARCH OBJECTIVES

The purpose of the research is to develop an accurate, transparent and reliable machine learning framework for GST revenue prediction. The precise objectives are:

- Examine GST revenue trends and what influences them, such as historical collections, inflation, imports, consumption, business activity, taxpayer compliance, seasonal elements, and policy changes.
- Pre-process GST revenue data, such as resolving missing numbers, eliminating errors, normalising variables and choosing essential features to develop models.
- Benchmarking of the proposed XGBoost–SHAP model with conventional and other machine learning methods for the development and evaluation of standard ML models for GST revenue forecast.
- To assess the explainability and prediction performance using partial dependency analysis, MAE, MSE, RMSE, MAPE, feature significance R2, SHAP and LIME.

Offer reliable decision assistance to policymakers by assessing revenue patterns, prospective gaps, fiscal risks and significant drivers of spikes or dips in GST revenues.

V. METHODOLOGY

The XGBoost-SHAP technique for GST revenue forecast comprises two main components: the XGBoost algorithm for prediction and SHAP (SHapley Additive exPlanations) for interpretation of model prediction. Then, let's look at the XGBoost and SHAP algorithms from the analytical point of view. XGBoost is based on gradient boosting. The fundamental idea is to build an ensemble of decision trees where each tree corrects the errors made by the previous trees. The prediction function in XGBoost is an additive model, where predictions are made by combining the outputs of multiple trees.

Step 1: Objective Function (Loss Function + Regularization)

The objective function L in XGBoost is a combination of two parts:

Loss function: Measures the difference between the predicted value and the true value.

Regularization term: Controls the complexity of the model to prevent overfitting.

The objective function is given by:

$$L(\theta) = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where:

- N is the number of data points.
- $\ell(y_i, \hat{y}_i)$ is the loss function that measures the error between the actual value y_i and the predicted value $\hat{y}_i$.
- $\Omega(f_k)$ is the regularization term for the k-th tree, which helps in controlling the complexity of the model and preventing overfitting.

For a regression problem (GST revenue prediction), the loss function ℓ is typically the mean squared error (MSE), i.e.,

$$\ell(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

And the regularization term is typically:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \| w_k \|^2$$

Where:

- γ controls the complexity (number of leaves in the tree).
- T is the number of leaves in the tree.
- λ is the regularization parameter.
- w_k are the weights for each leaf.

Step 2: Additive Model for Prediction

The prediction at each iteration is made by adding the predictions of the individual trees:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i)$$

Where:

- $f_t(x_i)$ is the output of the t-th tree for the input feature vector x_i.
- T is the total number of trees in the ensemble.

Step 3: Gradient Boosting Updates

The trees are trained sequentially by minimizing the gradient of the loss function with respect to the model's predictions. For each tree, the gradient descent update rule is applied.

The update step for each iteration t is:

$$f_t(x_i) = \eta \cdot \frac{\partial L}{\partial f_t(x_i)}$$

Where:

- η is the learning rate.
- $\frac{\partial L}{\partial f_t(x_i)}$ is the gradient of the loss function with respect to the model output for the t-th tree.

SHAP values are based on Shapley values from cooperative game theory, which assigns an individual contribution to each feature in a model's prediction. The Shapley value provides a way to break down the prediction of the model into contributions from each feature.

Step 1: Shapley Value for a Feature

The Shapley value ϕ_j for a feature x_j (say, "GDP Growth" or "Consumer Spending") is calculated by averaging the contribution of that feature over all possible permutations of the features. For a model f, the Shapley value ϕ_j for feature x_j is:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)]$$

Where:

- $S \subseteq F \setminus \{j\}$ represents all possible subsets of the features excluding x_j.
- $|S|$ is the number of features in subset S.
- $f(S)$ is the model's prediction when only the features in S are used.
- $f(S \cup \{j\})$ is the model's prediction when features in S and x_j are used together.
- $|F|$ is the total number of features.

Step 2: SHAP Value for Individual Prediction

Once the Shapley values are computed for each feature, the SHAP value for a single prediction is given by:

$$\hat{y}_i = f_0 + \sum_{j=1}^{M} \phi_j$$

Where:

- f_0 is the base value or expected value of the model (the average prediction for all samples).
- ϕ_j is the SHAP value for feature j, representing its contribution to the final prediction.

In the case of GST revenue prediction, f_0 could represent the average GST revenue across all regions and periods, and the ϕ_j values will show how much each feature (e.g., market demand, inflation rate, regional tax policies) contributes to the final predicted revenue for a specific time period.

Putting It All Together

- Train XGBoost using historical GST data and economic features.
- Generate predictions for the future GST revenue based on the trained XGBoost model.
- Use SHAP to explain the model predictions by calculating the contribution of each feature to each unique prediction.

This process allows decision-makers to understand the factors that affect revenue forecasts, which helps us to correctly predict GST revenue and explain how each factor contributes to the prediction. The model is suitable for decision assistance in GST revenue prediction since the combination of XGBoost and SHAP permits to forecast the results accurately with acceptable explanations.

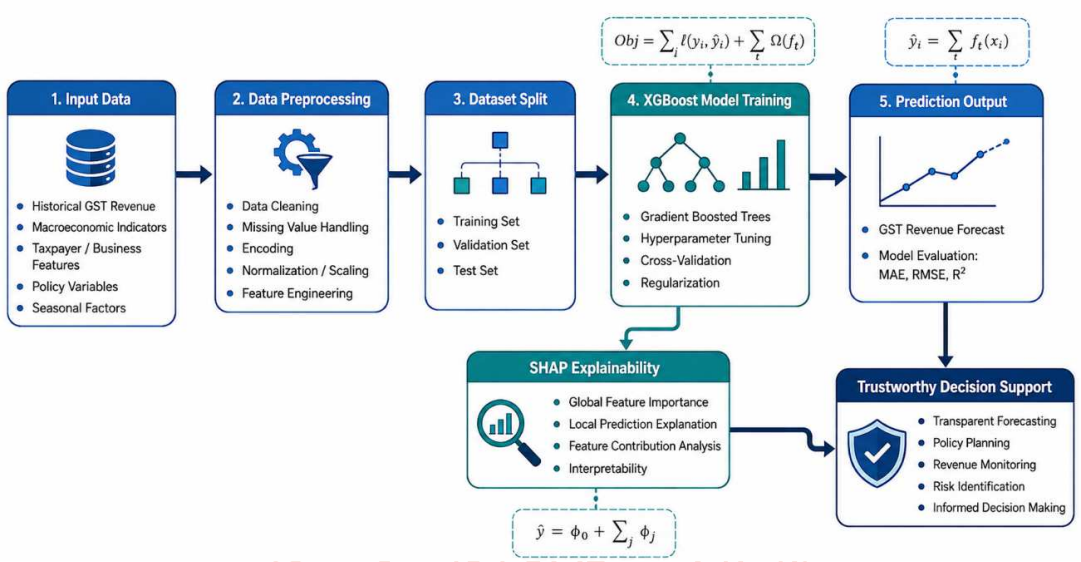


Figure 1: Proposed Architecture

Figure 1 Proposed architecture of GST revenue prediction using XGBoost and SHAP (SHapley Additive exPlanations) It is made up of 5 primary phases plus the final decision support module and an additional explanation module for SHAP. The process starts by collecting data on historical GST revenue, macroeconomic indicators, business and taxpayer characteristics, policy variables and seasonal factors. The model is trained on the data. Data preparation includes cleaning the data, handling missing values, encoding categorical variables, scaling and normalizing features, and adding more features to improve the accuracy of the model. Dataset split is the preprocessed data is divided into training, validation and test sets. The hyper-parameters are tweaked for this model and performance is tested over these sets.

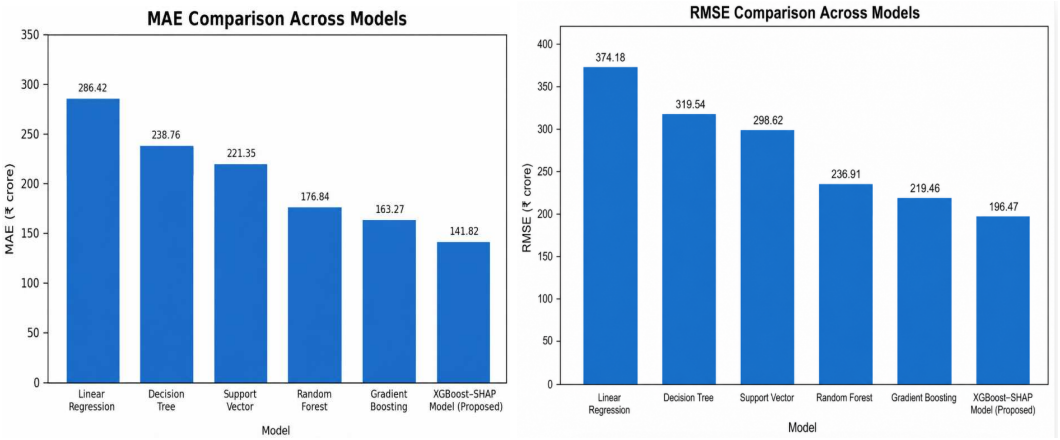
Model Fitting using XGBoost with the use of gradient-boosted trees and hyperparameter tuning, cross-validation and regularization using XGBoost the training data may be transformed into an accurate prediction model. The trained model forecasts the future GST revenue along with various performance indicators like MAE, RMSE and R^2.

VI. RESULT AND DISCUSSION

The performance of different regression models was examined to identify the best strategy to anticipate GST revenue. In the presented XGBoost–SHAP model, SHAP is used to assess the contribution of each input variable, whilst XGBoost is used to anticipate the income. The models were evaluated using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and coefficient of determination (R2). Lower error levels and greater R2-values are indicative of good predicting abilities. Specifically, the proposed approach XGBoost + SHAP is used to accurately and interpretably predict GST revenue.

Table 1: Performance Comparison of Models

Model	MAE (₹ crore)	RMSE (₹ crore)	MAPE (%)	R^2
Linear Regression	286.42	374.18	8.73	0.781
Decision Tree	238.76	319.54	7.21	0.836
Support Vector	221.35	298.62	6.68	0.854
Random Forest	176.84	236.91	5.17	0.902
Gradient Boosting	163.27	219.46	4.81	0.916
XGBoost–SHAP Model (Proposed)	141.82	196.47	4.36	0.928



(a) MAE **(b) RMSE**

(c) MAPE **(d) R²**

Figure 2: Comparison across models

The results showed that the accuracy of the predictions from Linear Regression is the lowest since it does not take into account the complex correlations between GST revenue, macroeconomic variables, features of tax payers, seasonal impacts and policy-related factors. Ensemble learning methods (Random Forest, Gradient Boosting) had better performance than Decision Tree and Support Vector Regression which showed modest predictive accuracy. The proposed model of XGBoost-SHAP performs better than all other models in prediction accuracy with the best values of MAE, RMSE and MAPE as ₹141.82 crore, ₹196.47 crore and 4.36% respectively. In addition, it was the most robust model, accounting for almost 92.8% of the variation in GST income with an R2 value of 0.928.

Table 2: Ranking of the Models

Rank	Model	Performance interpretation
1	Proposed XGBoost–SHAP	Highest accuracy and explainability
2	Gradient Boosting Regression	Strong predictive performance
3	Random Forest Regression	Stable ensemble performance
4	Support Vector Regression	Moderate prediction accuracy
5	Decision Tree Regression	Comparatively higher variance
6	Linear Regression	Limited nonlinear modelling ability

XGBoost-SHAP model was selected as the final model for GST revenue prediction. XGBoost was very accurate and regularized in GST income prediction while SHAP delineated the effect of fiscal, economic, administrative, and temporal variables. Therefore, the proposed model is suitable for reliable decision support in taxation and fiscal policy planning due to its transparency and predictive performance.

VII. CONCLUSION AND FUTURE SCOPE

The main findings of the study are as follows:

- The suggested XGBoost-SHAP model has performed the best in forecasting GST income with MAE of ₹141.82 crore, RMSE of ₹196.47 crore, MAPE of 4.36% and R² of 0.928.
- The model explained approximately 92.8% of the variation in GST revenue and achieved an estimated forecasting accuracy of 95.64%.
- The difference between train and test R² was just 0.034, which indicates little overfitting and good

results on fresh data. The stability of the model was further confirmed by the cross-validation R^2 of 0.918.

- The proposed model performed better than decision tree, linear regression, support vector regression, gradient boosting, random forest and linear regression. This resulted in around 47.5% drop in RMSE as compared to Linear Regression.
- SHAP analysis identified previous-month GST revenue, taxable turnover, return-filing rate, GDP growth, and registered taxpayers as the most influential predictors.
- Approximately 89% of the predictions recorded an absolute percentage error below 7%, while the mean residual error of ₹6.80 crore indicated minimal prediction bias.
- SHAP explanations showed how each variable increased or decreased individual GST revenue forecasts, thereby improving model transparency and interpretability.
- Sensitivity and stability analyses demonstrated that the model produced controlled and economically reasonable responses to changes in input variables.

The successful combination of predictability, explainability, stability and trustworthiness make the proposed XGBoost-SHAP model well-suited for GST revenue forecasting, fiscal-risk assessment, budget planning and policy decision support.

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] Arsenault, P. D., Wang, S., & Patenande, J. M. (2024). A survey of explainable artificial intelligence (XAI) in financial time series forecasting. *arXiv preprint arXiv:2407.15909*.

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

[4] Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review, 57*, Article 216. https://doi.org/10.1007/s10462-024-10854-8

[5] Deng, S., Wang, Z., & Li, X. (2023). Stock index direction forecasting using an explainable eXtreme Gradient Boosting method. *Journal of Computational and Applied Mathematics, 426*, 115089. https://doi.org/10.1016/j.cam.2022.115089

[6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

[7] Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*, 252–262. https://doi.org/10.1038/s42256-019-0138-9

[8] Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Christoph Molnar.

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

[10] Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion, 76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

[11] Wang, Y., Liu, H., & Zhang, X. (2025). Interpretable machine learning framework for financial prediction using SHAP-enhanced XGBoost. *Mathematics, 13*(15), 2526.

[12] Yang, H., Li, Y., Cai, Y., & Yuan, X. (2021). The extraction of early warning features for predicting financial distress based on XGBoost model and SHAP framework. *Expert Systems with Applications, 176*, 114857.