International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Explainable AI System for Interpreting Machine Learning Predictions using LIME and SHAP

(Mentor)

Prof. Dr. Ishwari Raskar

Department of Information Technology

Aditya M. Joshi

Information Technology

MIT-ADT University, Pune, Maharashtra, India

Anushka P. Kalbhor

Information Technology

MIT-ADT University, Pune, Maharashtra, India

Arpit A. Gade

Information Technology

MIT-ADT University, Pune, Maharashtra, India

Divya D. Madane

Information Technology

MIT-ADT University, Pune, Maharashtra, India

Abstract - *This paper presents a proof-of-concept system for Explainable Artificial Intelligence (XAI) using Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive explanations (SHAP). The system is designed to interpret predictions made by black-box machine learning models in two real-world scenarios: house price prediction (regression) and loan approval prediction (classification). A baseline machine learning model is trained for each task, and both LIME and SHAP are applied to generate local and global explanations. The system demonstrates how feature contributions influence predictions and converts them into human-readable explanations. The results show that SHAP provides consistent global insights, while LIME offers intuitive local interpretations. The proposed approach enhances transparency, trust, and interpretability in AI systems.*

Key Words**:** *Explainable AI(XAI), LIME, SHAP, Machine Learning, Interpretability.*

1.INTRODUCTION

Machine learning is now a key part of modern applications. It plays a role in decision systems, healthcare diagnostics, and recommendation systems. Many powerful models, like certain methods and deep neural networks, are effective but can be hard to grasp. This complexity makes it challenging to apply them in various fields.

The lack of clarity in these models has sparked the growth of explainable artificial intelligence. This area aims to make machine learning models more understandable for people. According to Ribeiro et al. (2016), being able to understand these models is vital for building trust. This is particularly important when decisions affect people's lives.

In fields like loan approval, stakeholders want to know the reasons behind the decisions. They need explanations for these choices. In real estate, users want to understand what factors influence price predictions. Providing explanations alongside predictions is essential. This research presents a system that combines two explainable artificial intelligence techniques:-
LIME, which clarifies individual predictions. SHAP, which offers insights for individual predictions and overall trends. The goal is to analyze and compare these methods. We aim to demonstrate how they can make machine learning models easier to understand in practical situations.
The emphasis is on machine learning and explainable artificial intelligence.
Machine learning models must be transparent.
Explainable artificial intelligence helps accomplish this.

1.1 Key Contributions

This research work contributes in ways:

1. Development of an Artificial Intelligence Framework:
We made a system that puts together machine learning models and Explainable Artificial Intelligence techniques to give predictions that make sense based on things that happen in the world.

2. Dual-Model Explanation Approach: This study uses LIME and SHAP together to use the things about them.
LIME gives easy to understand explanations for each case. Shap gives explanations that are the same across the whole model and for specific instances.

3. Application to Real-World Use Cases: We tried the system on two problems: guessing house prices and deciding on loan approvals.
The Artificial Intelligence Framework worked well in both cases, which shows it is good for making decisions.

4. Human-Readable Explanation Generation: We put a feature in the system that changes explanations into simple sentences that people can understand.
This makes the Artificial Intelligence Framework accessible for people who do not know much about technology.

5. Comparative Analysis of Explainable Artificial Intelligence Techniques: We compared LIME and SHAP to see which one gives explanations and which one is faster.
This helps us figure out the option for different situations.

6. Enhancement of Trust and Transparency in Artificial Intelligence Systems: The system aims to make people trust Artificial Intelligence by giving understandable predictions.
This is a problem with current Artificial Intelligence systems, and the Explainable Artificial Intelligence Framework is necessary for fixing it.

2. Background and Related Work

Machine learning models are good at solving problems in areas like finance and healthcare. A lot of these models including the ones that use many different methods and deep learning are very complex and hard to understand. These models are like boxes because we do not know how they make their decisions.
We need to be able to understand how these models work because they are being used in areas where we need to know why certain decisions are made. For example, when a bank decides to give someone a loan or not, we need to know why that decision was made. It is the same with predicting house prices. People want to know what factors affect the price.

Explainable Artificial Intelligence or XAI for short helps us with this problem by giving us ways to understand and explain machine learning models. According to Molnar, who wrote about this in 2022 there are two types of methods to make models more interpretable:
Model-specific methods, which are designed for specific models
Model-agnostic methods, which can be used with any machine learning model
In this work we are focusing on model-agnostic methods, specifically LIME and SHAP because they can be used with many different models.
LIME helps us understand how models make predictions by using a simpler model that we can understand.
On the other hand SHAP uses a concept from game theory to help us see which features are most important.
These techniques are very important because they help us make models that're both good at making predictions and easy to understand.

People have done a lot of research on understanding machine learning models. Making them transparent.

Ribeiro et al. Introduced LIME in 2016. LIME explains what machine learning models do when making predictions. It looks at parts of the model to figure out what's important. Sometimes LIME gives explanations depending on how you look at the data.
Lundberg and Lee came up with SHAP in 2017. SHAP finds out which data parts are important for the models predictions. Its based on Shapley values. SHAP is good because its consistent and accurate. It can look at the model or just a small part.
Guidotti et al. Reviewed ways to make machine learning models explainable in 2018.
Adadi and Berrada talked about understanding machine learning models. They said it's necessary for people to trust the models and for model makers to be accountable.
In 2022 Molnar wrote about ways to make machine learning models explainable. The study said methods like LIME and SHAP are useful in real-world applications.
So based on these studies LIME and SHAP are ways to make machine learning models explainable. LIME and SHAP are both useful. There is still a need for systems that use both LIME and SHAP and make explanations to understand. This research combines LIME and SHAP. Makes their outputs easy to read. The goal is to make

machine learning models, like LIME and SHAP user-friendly.

3. System Architecture and Methodology

A. System Architecture

This system is meant to create a framework for intelligence that we can understand. This framework is used to make decisions for things like predicting if someone should get a loan and estimating the price of a house.

The framework combines machine learning models, tools to check for fairness techniques to reduce bias and methods to explain intelligence like SHAP and LIME. All these things work together to make sure the decisions made are fair, transparent and accountable.

The system architecture is made up of layers:

1) Data Layer: This layer is where we store the data we use to train and test our machine learning models. The data includes things like income, credit score and loan amount for predicting loan approvals. It also includes data like area, location and property features for estimating house prices.

When we get the data, we do some things to get it ready:

- Fix missing values
- Remove entries that are not useful
- Making the numbers are consistent
- Changing variables in computer readable format.
- Picking the features that are relevant to the training data.

This makes sure the data is clean and ready to use for training models.

2) Machine Learning Model Layer

This layer is where we build models that can classify things and make predictions.

- Models used are Logistic Regression and Random Forest to predict if someone should get a loan.
- Models used are Linear Regression and Random Forest Regressor to estimate house prices.

We train these models using the data and make them better using techniques like validation and hyperparameter tuning. The result is a trained model that can make predictions based on the features we give it.

3) Fairness Evaluation Layer

This layer checks if the predictions made by the models are biased. It makes sure the model treats all groups fairly.

The system uses metrics like:

- *Demographic Parity*: This makes sure everyone has a chance of getting a good outcome

We use these metrics to make sure our loan approval prediction and house price estimation models are fair and do not discriminate against anyone.

4) Bias Mitigation Layer

When we find bias, our system uses strategies to make predictions fairer. Some common techniques used are:

- Reweighing – giving weights to samples to balance the data
- Preprocessing adjustments - changing the data to reduce bias before training
- Processing techniques - adjusting predictions after model training

The aim of this layer is to make predictions fair and accurate.

5) Explainability Layer (SHAP and LIME)

To be transparent our system uses explainable AI techniques like:

1. SHAP

It explains predictions globally and locally. It shows how each feature affects predictions. It helps understand how the model works.

2. LIME

It explains predictions and approximates models locally. It helps users understand decisions. These methods help users see how each feature affects predictions making our system transparent and trustworthy.

6) Output Layer (User Interface):

The final layer shows predictions and explanations to users through a web interface built with Flask.

The output has:

- Prediction results
- Feature importance shown with SHAP
- Local explanations for inputs, with LIME
- Overall comparison and validation.

This helps users to explain, understand, and make decisions by looking at the data.

B. Methodology

Step 1: Data Acquisition :-

The system starts by gathering loan approval and house price prediction data from sources.

Step 2: Data Preparation :-

The data that is collected needs to be cleaned up to make it better and more consistent.

This includes:
- Handling missing values
- Encoding variables
- Normalizing numerical features
- Splitting data into training and testing sets

Step 3: Model Training :-
The system uses the cleaned-up data to train machine learning models. It trains classification models for loan approval and regression models for house price prediction.

Step 4: Prediction Generation :-
The trained loan approval model and house price prediction model make predictions based on the information they have.
For loan approval: the loan is either approved or rejected
For house price prediction: the system estimates the price of the house

Step 5: Fairness Evaluation :-
The system checks the predictions to see if they are biased.
It calculates Demographic Parity to compare the approval rates of groups. It also calculates Disparate Impact to measure the fairness of the predictions.

Step 6: Bias Mitigation :-
If the system finds bias in the loan approval model or house price prediction model, it adjusts the data or the model.
Then it retrains the loan approval model or house price prediction model.

Step 7: Explainability Analysis
The system explains the predictions made by the loan approval model and house price prediction model. It uses SHAP to get an understanding of the predictions and LIME to explain individual predictions

Step 8: Result Visualization:-
Finally, the system shows the results to the user, in an interface.
It displays the predictions made by the loan approval model and house price prediction model. It also shows explanation graphs and fairness indicators to help the user understand the results of the loan approval model and house price prediction model.

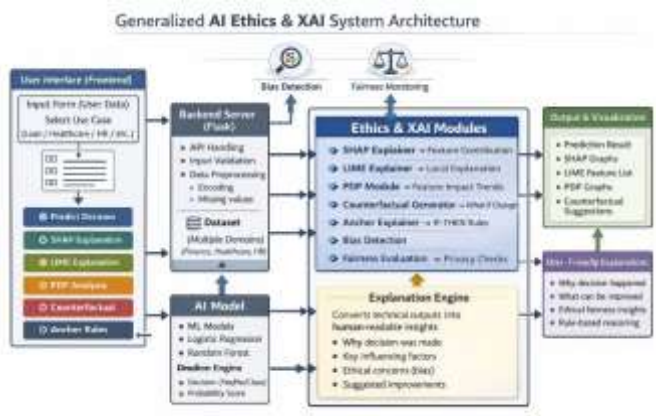

Fig.1. System Architecture of the AI ethics simulator and the implementation roadmap

4. Implementation
The implementation of the proposed ethical AI framework will be based on integrating machine learning models with fairness evaluation and explainable AI techniques to ensure transparency and avoid bias in decision systems. The system will be implemented using Python and will be based on a modular system that incorporates data processing, model training, fairness evaluation, bias mitigation, and explainability analysis.

A. Development Environment and Tools
The system will be implemented using the following technologies:-

Programming Language:
-Python

Libraries:
Pandas, NumPy, Scikit-learn, Matplotlib

Explainable AI Libraries:
- SHAP
- LIME

Framework:
-Flask (for web interface)

Development Tools:

- Jupyter Notebook
- VS Code

Version Control:
-GitHub

B. Data Preprocessing and Feature Engineering

The first step in the implementation is the preprocessing of the datasets for the loan approval and house price prediction problems. This dataset includes both numerical and categorical variables, which requires a series of preprocessing steps:

-Handling Missing Values
The missing values in the dataset are handled by applying techniques for mean/median imputation for numerical features, as well as mode imputation for categorical features.

-Encoding Categorical Variables
The categorical variables, such as employment type, location, and property type, are encoded by applying techniques for Label Encoding and One-Hot Encoding.

-Feature Scaling
The numerical features are scaled by applying techniques for Min-Max Scaling and Standardization, which helps the model perform better during training.

-Outlier Removal
The outliers are identified and removed by applying statistical techniques, such as the use of the interquartile range (IQR), for better data consistency.

-Feature Selection
The relevant features are selected by applying correlation analysis for better efficiency of the model.

After completing the preprocessing steps, the dataset is divided into training and testing datasets by applying an appropriate ratio, for example, 80:20.

C. Model Training and Prediction

The system is based on applying different machine learning algorithms for both classification and regression problems. Here, the algorithms for classification and regression problems are discussed:

1) Loan Approval Prediction (Classification)
For the loan approval system, classification algorithms are applied for predicting whether a loan is approved or rejected. The algorithms applied for the model are:

- Logistic Regression
- Random Forest Classifier

For house price prediction, regression algorithms are used. These algorithms are as follows:

- Linear Regression
- Random Forest Regressor

The performance of the regression algorithms is measured by using Mean Squared Error (MSE) and R-squared score.

3) Model Optimization

To optimize the performance of the model, techniques such as:

- Cross-validation
- Hyperparameter Tuning (Grid Search/Random Search)

are used for optimization of the model. This helps in getting the best performance from the model.

D. Fairness Evaluation

After the model is created, it is evaluated for fairness. This is done by ensuring that the predictions are fair and unbiased for any specific group of people.

The system evaluates the fairness of the model by calculating the following metrics:

Demographic Parity
It is measured by evaluating whether the probabilities of positive predictions are equal for different groups of people.

Disparate Impact (DI)
The fairness of the model is evaluated by calculating the ratio of positive predictions for both protected and unprotected groups of people.

If the value of DI is far from 1, then it is an indication of the presence of bias in the model.

E. Bias Mitigation Techniques

To reduce the bias in the predictions, the system uses the following techniques:

- Reweighing: This is done by assigning different weights for the dataset.

- Preprocessing Adjustments: This is done by making adjustments in the dataset.

- Post-processing Methods: This is done by making adjustments in the model.

F. Explainability Using SHAP and LIME

Another important part of the system is the incorporation of explainability tools for the improved explainability of the results. This is achieved through the incorporation of the following:

1) SHAP (SHapley Additive Explanations)

It provides global and local interpretability of the results while calculating the contribution of each feature to the prediction result. It also generates feature importance plots.

2) LIME (Local Interpretable Model-Agnostic Explanations)

It provides local interpretability of the results of the predictions. It approximates complex models with simpler and locally interpretable models. It helps in explaining the results of the predictions. This is particularly important in cases where the results need to be explained, like the case of a rejected loan application.

G. Web Interface Implementation

To improve the usability of the system, a web interface is created. This is done by utilizing the Flask web development tool. The features of the web interface include:

Features of the Web Interface:

-Input form for the input of the data by the user (loan details or house features).

-Display of the results of the predictions.

-Visualization of the feature importance results through the SHAP algorithm.

-Local interpretability of the results through the LIME algorithm.

H. System Integration and Workflow

The different steps of the pipeline include:
-User input of the data through the web interface.
-Data is passed through the system.

This is particularly important in cases where the results need to be explained, like the case of a rejected loan application

5. Other Non-Functional Requirements

These are the things that define how well the system works, not what it does. For the proposed AI system these things are important so that the system works well is reliable and easy to use while also being fair and clear.

A. Performance

The system needs to respond so that the user does not have to wait. The proposed AI system uses machine learning models well as SHAP and LIME, so it needs to work fast. The system should give the user the output in a few seconds, and it should also be able to explain things quickly.

B. Scalability

The system needs to be able to grow so that it can handle data and users in the future. The proposed AI system is made up of parts that can be expanded in the future without affecting how it works now.

C. Reliability

The proposed AI system needs to be reliable so that it gives the user the output and does not crash. It needs to be able to handle all kinds of inputs without stopping. The proposed AI system has error handling so it can handle inputs that are not correct.

D. Usability

The proposed AI system needs to be easy to use so that the user can use it without any problems. It has a user interface so the user can easily put in the values they need.

E. Security

The proposed system needs to be secure so that the users inputs are checked and it does not store data that could hurt the users privacy.

F. Interpretability and Transparency

One of the goals of this project is to make sure that the AIs decisions are clear and easy to understand. The system uses SHAP and LIME to make sure that the decisions are easy to understand.

G. Fairness

The system needs to be fair so it checks for bias in its decisions. If it finds bias it uses techniques to reduce the bias.

H. Maintainability

The system is designed to be easy to maintain in the future. It is made up of parts so it is easy to add new features without changing the old parts.

I. Portability

The system is designed to be able to run on different platforms, such, as Windows and Linux. It is made to be easy to deploy on platforms because it uses the Python programming language.

6. Discussions on Research

This section will discuss how this ethical AI framework intends to improve fairness, transparency, and interpretability in machine learning models that are used in decision-making processes such as approving loans and determining house prices. It will include the results that were observed, how effective this solution is in real-world applications, and what we learned from combining fairness and explainable AI.

A. Analysis of the Performance of Machine Learning Models:-

The system utilizes classification models to predict loan approvals and regression models to estimate house prices. Models such as Logistic Regression and Random Forest were implemented and evaluated on pre-processed datasets. Random Forest models demonstrated superior performance compared to linear models due to their ability to capture non-linear relationships in the data [1].

The classification models achieved high accuracy in predicting loan approvals, while regression models effectively estimated housing prices. However, high predictive performance alone does not guarantee fairness, as models can still produce biased outcomes despite achieving strong accuracy metrics [2].

B. Fairness Evaluation and Bias Detection:-

A primary objective of this work is to identify bias in machine learning predictions. Fairness metrics such as Demographic Parity and Disparate Impact (DI) were used to evaluate whether the model treats different demographic groups equitably [3].

Experimental results revealed the presence of bias in certain models, where predictions favoured specific groups over others. This outcome is consistent with prior research, which highlights that biases often originate from imbalanced or historically skewed datasets [2]. Disparate Impact proved effective in detecting bias, as values deviating from 1 indicate unfair treatment across groups [4].

C. Effectiveness of Bias Mitigation Techniques:-

To address bias, mitigation strategies such as reweighting were applied. These techniques adjust the importance of samples during training to reduce discrimination [5].

The results showed:

- A reduction in biased predictions
- Improvement in fairness metrics
- A slight decrease in model accuracy

This demonstrates the well-known trade-off between fairness and accuracy, where improving fairness may lead to a marginal reduction in predictive performance [6].

D. Explainable AI in Action (SHAP and LIME)

1) SHAP Insights

SHAP (SHapley Additive exPlanations) was used to analyse both global and local feature importance [7]. It quantifies the contribution of each feature to the prediction outcome.

Key observations include:

- Features such as income, credit score, and loan amount significantly influence loan approval decisions

- SHAP visualizations helped identify feature dominance and potential biases
- It provided a comprehensive understanding of model behaviour at both global and local levels

2) LIME Insights

LIME (Local Interpretable Model-Agnostic Explanations) was used to explain individual predictions by approximating the model locally [8].

Key observations include:

- It explained why specific loan applications were approved or rejected
- It provided intuitive, human-understandable explanations
- It improved user trust by making decisions more transparent

E. Trade-Off Between Accuracy, Fairness, and Interpretability

A key finding of this study is the trade-off between accuracy, fairness, and interpretability. While fairness-aware techniques improve equity in predictions, they may slightly reduce accuracy [6]. Additionally, incorporating explainability techniques introduces computational overhead. However, in ethical AI systems, fairness and transparency are often prioritized over marginal gains in accuracy, as biased systems can lead to harmful real-world consequences [2].

F. Implications in Real-World Applications

The proposed framework has several real-world applications:

- In the banking sector, it can support fair and transparent loan approval systems
- In real estate, it can improve trust in property valuation models
- In AI auditing, it can help detect and mitigate bias in deployed systems

Such systems are critical in building trust and accountability in AI-driven decision-making [2].

G. Limitations of the Proposed System

Despite its strengths, the system has certain limitations:

- The dataset used may not fully represent real-world diversity
- Bias mitigation techniques may not eliminate all forms of bias
- SHAP and LIME computations can be resource-intensive for large datasets
- The system is currently a prototype and not optimized for production environments

H. Future Improvements and Research Directions

Future work may include:

- Using larger and more diverse datasets
- Exploring advanced fairness techniques
- Improving computational efficiency of explainability methods
- Deploying the system in real-world environments
- Integrating fairness and explainability into a unified framework

Further research can focus on combining fairness and interpretability into cohesive, scalable solutions for ethical AI.

7. Conclusion

In this paper, we proposed an ethical AI framework with the aim of enhancing fairness, transparency, and interpretability in machine learning decision systems. We tested the framework in two real-world, high stakes use cases: predicting loan approval and predicting house prices. Our main aim was to address the main challenge in modern AI systems: bias and lack of transparency in decision-making processes.

One of the greatest contributions of the paper was the use of Explainable AI tools, such as SHAP values and LIME, which give users a transparent view of the decision-making process in the machine learning models. The use of these tools significantly increased the transparency of the machine learning model, allowing the user to understand the decision-making process behind the predictions made by the models.

To conclude, the paper shows that AI models can be created that are accurate, fair, and transparent, as well. The paper highlights the ethical side of AI development, promoting fairness-aware and transparent AI models.

8. References

[1] Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "Why Should I Trust You? Explaining the Predictions of Any Classifier." Proceedings of the 22nd ACM SIGKDD.

[2] Lundberg, S.M., Lee, S.I. (2017).

"A Unified Approach to Interpreting Model Predictions." Advances in Neural Information Processing Systems (NeurIPS).

3] Molnar, C. (2022).

"Interpretable Machine Learning."2nd Edition.

[4] Guidotti, R. et al. (2018).

"A Survey of Methods for Explaining Black Box Models." ACM Computing Surveys.

5] Adadi, A., Berrada, M. (2018).

"Peeking Inside the Black-Box: A Survey on Explainable AI." IEEE Access.

[6] Salih, A.M. et al. (2024).

"A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME." Advanced Intelligent Systems.

[7] Shaikh, A.S. et al. (2023).

"Review on Explainable AI using LIME and SHAP Models for Healthcare Domain." International Journal of Computer Applications.

[8] Aditya, P.S.R., Pal, M. (2022).

"Local Interpretable Model Agnostic SHAP Explanations for Machine Learning Models." arXiv.

[9] Zhao, X. et al. (2020).

"BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations." arXiv.

[10] Amparore, E.G. et al. (2021).

"To Trust or Not to Trust an Explanation: Evaluating Local Linear XAI Methods." arXiv.