Predicting Loan Approval with Interpretable Machine Learning and XAI Techniques

Manikanta Nagulapalli(6361389)

Abstract—**This paper discusses about the use of interpretable machine learning (ML) and explainable artificial intelligence (XAI) techniques to predict loan approvals. The goal is to develop a model that not only performs well in terms of accuracy but is also interpretable and transparent in its decision-making process. The dataset used in this study includes various factors that affect loan approvals, such as credit history, employment history, loan amount, and loan term. Using of different ML algorithms and XAI techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) to explain the predictions made by the model. LIME generates local explanations for individual predictions, while SHAP provides a global explanation for the entire model. To evaluate the performance of the model using accuracy. Results show that the Random Forest algorithm performs better than other algorithms in terms of interpretability and accuracy. Additionally, the XAI techniques provide valuable insights into the model's decision-making process and help improve transparency and trust in the predictions made by the model. Overall, this study demonstrates the importance of using interpretable ML and XAI techniques in loan approval predictions. The combination of interpretable ML algorithms and XAI techniques can provide valuable insights into the model's decision-making process, improve transparency, and increase trust in the model's predictions.**

Index Terms—**component, formatting, style, styling, insert**

I. Introduction

The loan approval procedure is critical in the financial business since it includes a substantial amount of money and risk. Accurate loan approval prediction is critical for both lenders and borrowers, since it impacts both sides' financial well-being. Manual study of different aspects such as credit history, work history, loan amount, and loan term is used in traditional loan approval methods. However, as the volume of loan applications grows and the demand for speedier decision-making grows, there is a rising interest in applying machine learning (ML) and artificial intelligence (AI) approaches to forecast loan approvals.

In the realm of loan approval forecasts, interpretable ML and explainable artificial intelligence (XAI) algorithms have received a lot of interest in recent years. The objective of employing interpretable ML and XAI approaches is to produce models that are not only accurate but also interpretable and transparent in their decision-making process. We may obtain insights into the model's decision-making process and increase transparency and trust in the model's predictions by applying interpretable ML and XAI approaches.

Artificial intelligence (AI) algorithms, particularly those based on deep neural networks, have transformed the way we approach real-world jobs traditionally performed by people. There has been a substantial surge in the application of Machine Learning (ML) algorithms to automate different aspects of scientific, business, and social processes in recent years. This increase can be traced in part to the expansion of research in the field of machine learning (ML), known as Deep Learning (DL), in which hundreds (even billions) of neural parameters are taught to generalize and execute particular tasks. DL algorithms have been successfully applied in healthcare, ophthalmology, developmental diseases, autonomous robotics, and transportation, Image processing categorization and identification, voice and audio processing, and cyber-security are just a few examples of how DL algorithms are used in our daily life. Access to high-performance compute nodes via cloud computing ecosystems, high-throughput AI accelerators to improve performance, and large-scale datasets and storage enable deep learning providers to research, test, and operate ML algorithms at scale on small edge devices, smartphones, and AI-based web-services that use APIs for wider exposure to any applications.

Deep Neural Networks (DNNs) are difficult to grasp and interpret due to their enormous number of parameters. Regardless of cross-validation accuracy or other assessment criteria that may suggest high learning performance, DL models may learn or fail to learn representations from input that a human would consider important. Explaining DNN judgments needs understanding of DNN internal processes, which is frequently lacking among non-AI professionals and end-users focused on finding accurate solutions. As a result, the capacity to comprehend AI judgments is sometimes seen as secondary in the quest to attain cutting-edge outcomes or outperform human-level accuracy.

Recently, there has been a surge in interest in XAI, including from governments, particularly with the European General Data Protection Regulation (GDPR), which highlights the relevance of ethics, trust, bias, and the influence of adversarial cases on AI. Miller et al. stated that one of the key reasons people seek explanations for certain actions is curiosity. Another motive might be to improve learning by iterating model design and producing better outcomes. Each explanation should be consistent across similar data points and produce consistent or similar explanations on the same data point throughout time.

Explanations should make the AI algorithm more expressive in order to increase human comprehension, decision-making confidence, and to encourage unbiased and just judgments. To ensure openness, trust, and fairness in the ML decision-making process, ML systems must provide an explanation or

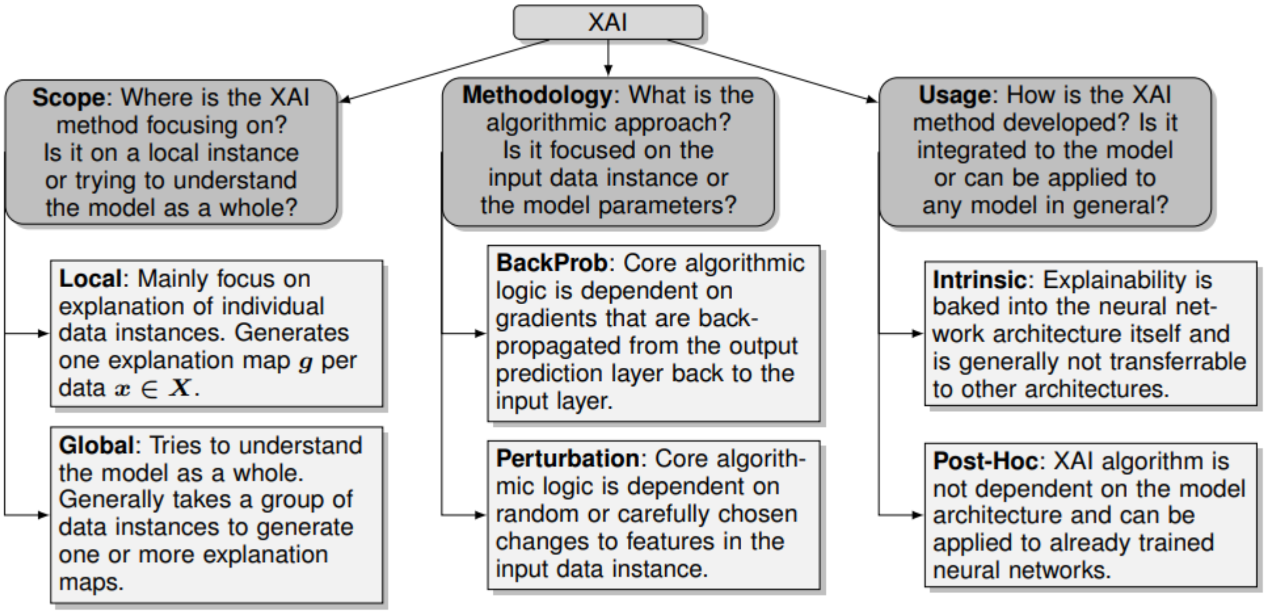

XAI
Scope: Where is the XAI method focusing on? Is it on a local instance or trying to understand the model as a whole?
Local: Mainly focus on explanation of individual data instances. Generates one explanation map $\boldsymbol{g}$ per data $\boldsymbol{x} \in \boldsymbol{X}$.
Global: Tries to understand the model as a whole. Generally takes a group of data instances to generate one or more explanation maps.
Methodology: What is the algorithmic approach? Is it focused on the input data instance or the model parameters?
BackProb: Core algorithmic logic is dependent on gradients that are back-propagated from the output prediction layer back to the input layer.
Perturbation: Core algorithmic logic is dependent on random or carefully chosen changes to features in the input data instance.
Usage: How is the XAI method developed? Is it integrated to the model or can be applied to any model in general?
Intrinsic: Explainability is baked into the neural network architecture itself and is generally not transferrable to other architectures.
Post-Hoc: XAI algorithm is not dependent on the model architecture and can be applied to already trained neural networks.

Fig. 1. XAI (Explainable Artificial Intelligence)

an interpretable answer.

An explanation is a method of validating an AI agent's or algorithm's output decision. An explanation for a cancer detection model employing microscopic pictures might be a map of input pixels that contribute to the model output. An explanation for a voice recognition model might be the power spectrum information at a certain moment that contributed more to the current output decision.

II. Data Set

A loan approval dataset is a collection of information about individuals who have applied for loans from financial institutions. This dataset contains various features about the applicant, such as their personal details, employment status, credit history, financial information, loan amount requested, and the final decision whether the loan was approved or not.

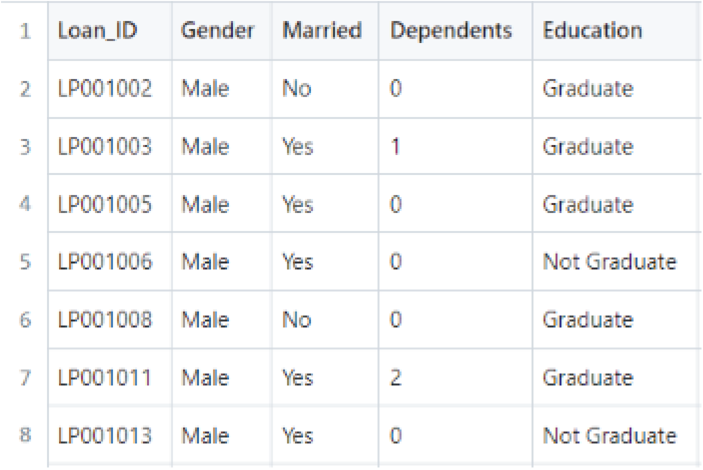
The purpose of this dataset is to help financial institutions make informed decisions about whether to approve or reject loan applications. By training machine learning models with this dataset, financial institutions can predict the likelihood of a loan being approved based on various features. This can help automate the loan approval process and make it more efficient.

Loan approval datasets usually contain a mix of numerical and categorical features. Numerical features may include income, credit score, and loan amount requested. Categorical features may include employment status, gender, and marital status. The dataset may also contain missing values or outliers, which must be handled appropriately before building a model.

It's important to note that loan approval datasets may contain sensitive personal information about individuals. Therefore, any analysis of such datasets must be done in a way that protects individuals' privacy and conforms to applicable laws and regulations. Overall, loan approval datasets can be a valuable tool for financial institutions to improve their loan approval process and make more informed decisions. The goal of this project is to use a publicly available dataset to predict whether a person will opt for a bank loan or not. The dataset contains 13 input columns and one binary output label that indicates whether a loan has been taken or not. To eliminate irrelevant features, correlation analysis is used to remove columns that are not correlated with the output variable. After this, the dataset is left with 10 input features.

The dataset contains 5000 samples, with 480 of them having a label of 1 (loan is taken). The dataset is split into training and testing sets in a 7:3 ratio. For the demonstration of a decentralized setup, the training dataset is further split into two parts: personal details (6 features) and bank-specific details (4 features).

The personal details part includes information such as education level, number of family members, annual income, average credit card spending, years of work experience, and the value of the house mortgage. The bank-specific details part includes information such as whether the customer has a security account with the bank, whether they have a CD account with the bank, whether they have a credit card issued by the Universal Bank, and a combination feature.



1	Loan_ID	Gender	Married	Dependents	Education
2	LP001002	Male	No	0	Graduate
3	LP001003	Male	Yes	1	Graduate
4	LP001005	Male	Yes	0	Graduate
5	LP001006	Male	Yes	0	Not Graduate
6	LP001008	Male	No	0	Graduate
7	LP001011	Male	Yes	2	Graduate
8	LP001013	Male	Yes	0	Not Graduate

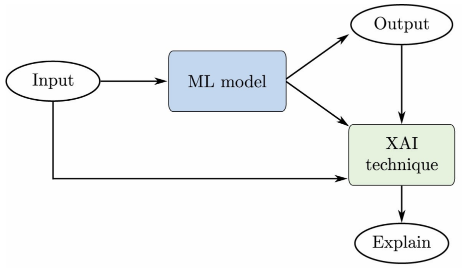
Fig. 2. Sample data structure personal Information

III. Methodology

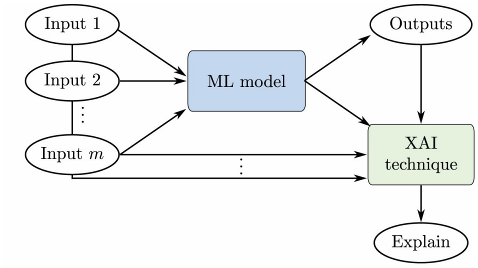
Due to its many benefits, random forest is regarded as one of the top machine learning algorithms for loan prediction. First, it can easily manage missing data, which is a regular occurrence in loan prediction datasets. This makes it more likely that the model will be able to make reliable predictions even in the absence of certain input data. Second, by combining the findings from many decision trees, Random forest is recognized to lessen overfitting. This ensures that the model operates effectively on untested data and enhances the generalization performance of the model.

A local explanation explains how a model arrived at a certain conclusion for a single instance or observation. It aids in comprehending which characteristics of the input data had a substantial influence on the model's conclusion. By creating explanations for each prediction, methods like LIME and SHAP offer local explanations.

On the other hand, a global explanation shows how a machine learning model behaves generally across all data points. It aids in determining which characteristics significantly influence how the model makes decisions. It helps us to comprehend the model's advantages and disadvantages and gives us a glimpse into how the model functions within. Global explanations can be provided by methods like partial dependency plots and the relevance of permutation features.



(a) Local explanation



(b) Global explanation

Understanding machine learning models requires both local and global explanations, and they have different functions. While a global explanation is valuable for understanding the model's general behavior and pinpointing areas for development, a local explanation is useful for understanding the thinking behind a single prediction.

Additionally, Random Forest can handle both category and numerical data, making it suitable for loan prediction when the input data may contain characteristics like job status, educational attainment, and credit score, which can be both categorical and numerical. It also offers a feature importance rating, which aids in determining which features are crucial for the loan prediction assignment. Both feature selection and data exploration may benefit from this.

Random forest is renowned for its excellent accuracy and resistance to data noise, and it is highly resilient to outliers, which might occur often in loan prediction datasets. These elements make it an excellent option for loan prediction, where the model's accuracy is essential for making judgments on loan approval that are well-informed. In conclusion, Random forest is a flexible and strong algorithm that can handle many kinds of data and provide high accuracy in loan prediction applications.

A. *Logistic Regression*

For binary classification problems, where the objective is to predict one of two potential outcomes based on the input features, the machine learning approach known as logistic regression is utilized. It operates by simulating the relationship between the input features and the likelihood of a favorable result.

The algorithm converts the input information to a probability between 0 and 1 using a logistic function. As the input value approaches negative and positive infinity, respectively, the logistic function asymptotically approaches 0 and 1 with a distinctive S-shaped curve.

Maximum likelihood estimation, which determines the values of the parameters that maximize the probability of the observed data, is used by the logistic regression method to estimate the parameters of the logistic function.

B. *Decision Algorithm*

The decision algorithm is a machine learning algorithm used for classification tasks, similar to the decision tree algorithm. It works by constructing a model that divides the input space into rectangular regions, each of which is associated with a specific class label.

The algorithm selects the best feature and threshold to split the data based on a criterion such as information gain or Gini impurity. It then recursively applies this process to each of the resulting subspaces until a stopping condition is met, such as a minimum number of samples per leaf or a maximum depth of the tree.

1) LIME (Local Interpretable Model-Agnostic Explanations): LIME (Local Interpretable Model-Agnostic Explanations) is a machine learning model interpretability approach. It is intended to provide context for a model's individual instance predictions. LIME is a technique that may be used with any machine learning model since it is model-agnostic.

LIME operates by roughly simulating the behavior of the model close to a specific instance. In the neighborhood of the instance being discussed, it creates a local linear model that simulates the behavior of the sophisticated machine learning model. The local linear model can assist shed light on how the complicated model generates its predictions since it is simple to comprehend.

In order to fit a linear model that roughly approximates the behavior of the complicated model, LIME first generates a series of perturbed copies of the instance being described. By making minor, comprehensible changes to the original instance, perturbed instances are created. The perturbed instances, for instance, may be produced by masking off specific picture pixels for an image classification job.

By analyzing how closely the local linear model produced by LIME resembles the complex model's behavior on the perturbed instances, its quality may be determined. The weights that the linear model gave the input features show how much each feature contributed to the final forecast for the case under investigation.

2) SHAP (SHapley Additive exPlanations): A approach for describing the predictions of machine learning models called SHAP (SHapley Additive exPlanations) is model-independent. It is based on the cooperative game theory idea of Shapley values, which is utilized to calculate how much each characteristic contributes to the prediction.

A unifying framework for analyzing the predictions of any machine learning model, including black-box models, is provided by the SHAP technique. Due to its limited nature, it can only explain the prediction of one occurrence at a time. For each feature in the dataset, the SHAP technique creates a collection of Shapley values that indicate how much of a contribution that feature made to the prediction. The average deviation in the prediction when a feature is included and omitted from all potential subsets of features is known as the feature's Shapley value.

The Shapley plot for each instance is then produced using the SHAP method using these Shapley values. The Shapley plot shows how each attribute contributed to the prediction for one specific case. While negative values show that the characteristic diminishes the prediction, positive values show that the feature boosts it.

The Shapley plot is an effective visualization technique that enables viewers to quickly comprehend how each variable contributes to the forecast. Users may determine which characteristics are most crucial for a certain prediction and how they affect the forecast by looking at the Shapley plot.

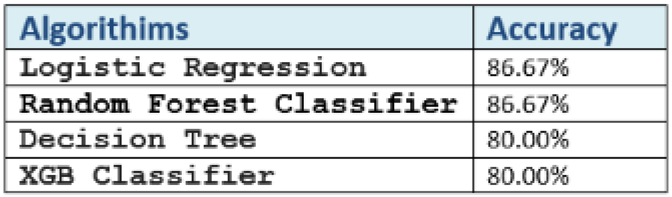
The SHAP technique also offers global feature significance metrics, which show the average contribution of each feature to the prediction across all instances in the dataset, in addition to the Shapley plot.

All things considered, SHAP is a potent and well-liked approach for elucidating the predictions of machine learning models. It is especially useful for evaluating complicated models and determining which aspects are most crucial for a certain prediction since it may offer model-independent, local explanations.

IV. Results

The figure uses the hues red and blue. Blue characteristics show unfavorable contributions to loan denial, while red features show good contributions to loan approval.

TABLE I
Comparing Accuracy of ML Algorithim



Algorithims	Accuracy
Logistic Regression	86.67%
Random Forest Classifier	86.67%
Decision Tree	80.00%
XGB Classifier	80.00%

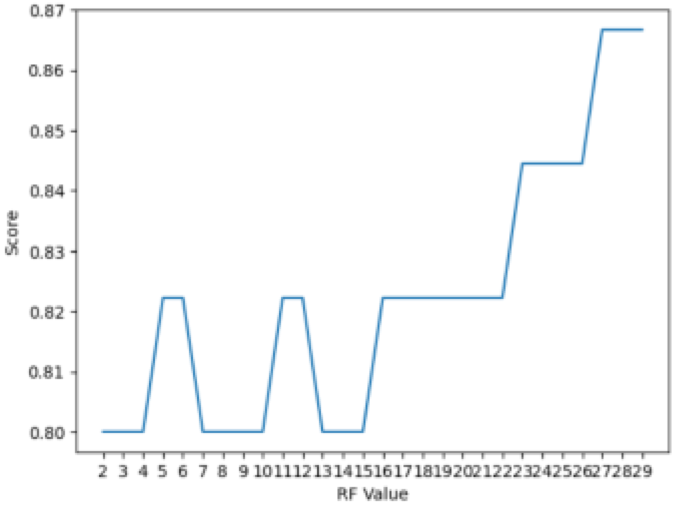


Fig. 3. Cell output features

Different machine learning algorithms may be used to forecast the chance that a loan will be accepted or denied

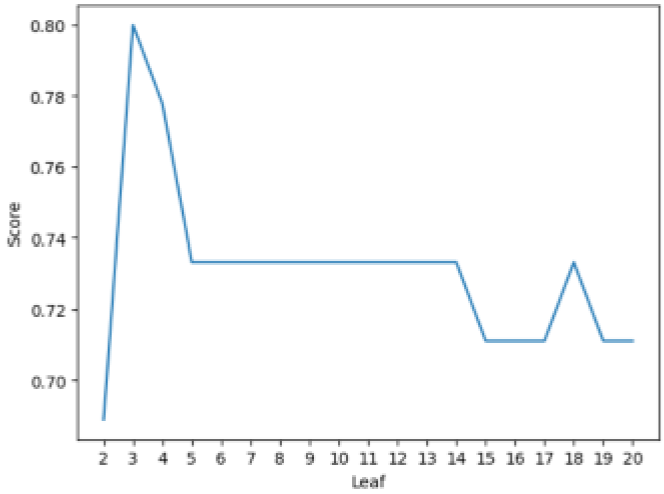


Fig. 4. Cell output features

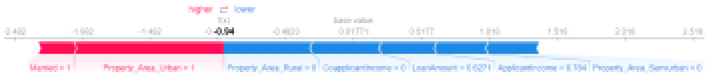


Fig. 5. Cell output features

when it comes to loans. In this instance, the loan approval status was predicted using four algorithms: logistic regression, random forest classifier, decision tree, and XGB classifier.

According to the findings, random forest classifier and logistic regression both had an accuracy of 86.67

With an accuracy of 80.00

Overall, the findings imply that the loan prediction algorithms logistic regression and random forest classifier are promising. To maintain honest and open lending procedures, it's crucial to dig deeper into and analyze the factors that have the biggest influence on the decision to approve a loan.

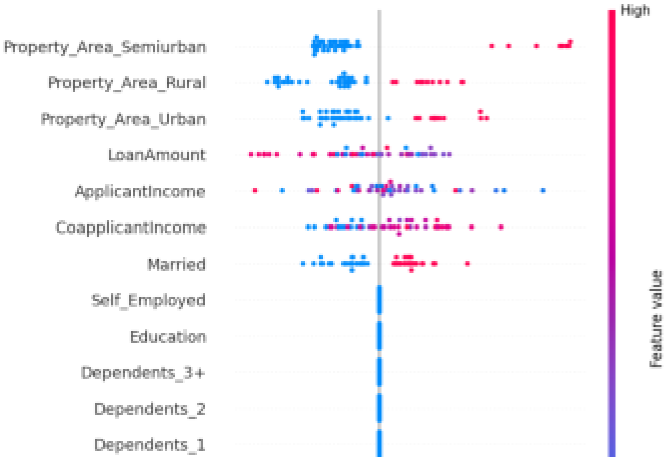


Fig. 6. SHAP value(Imapct Model Output)

The SHAP values, where positive values reflect positive contributions towards loan approval and negative values represent negative contributions, may also be used to provide an overview of the overall feature relevance. Each attribute is shown on the y-axis, with red denoting high values and blue denoting low values. By examining this story, we can see that the applicant's married status has always had a beneficial influence on loan acceptance, however a large loan amount has a negative influence. Analyzing the effect of each item on loan approval requires reading the color indicators and SHAP values.

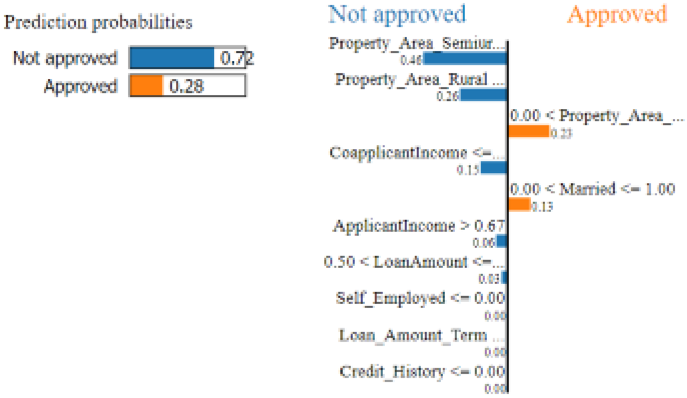
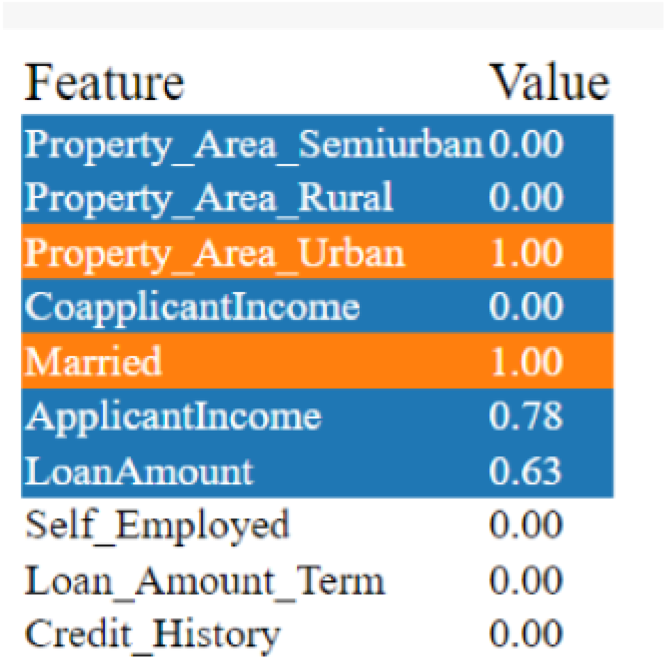


Fig. 7. Random Forest accuracy



Feature	Value
Property_Area_Semiurban	0.00
Property_Area_Rural	0.00
Property_Area_Urban	1.00
CoapplicantIncome	0.00
Married	1.00
ApplicantIncome	0.78
LoanAmount	0.63
Self_Employed	0.00
Loan_Amount_Term	0.00
Credit_History	0.00

Fig. 8. Decision Tree accuracy

LIME also follows a similar process as SHAP. However, before applying LIME, we need to initialize a tabular explainer, which has a specific explainer for each type of data such as text, tabular, images, etc. The LIME explainer takes four parameters, which include training data, feature names, classification.

Conclusion

In conclusion, it is critical to anticipate loan approval utilizing interpretable machine learning and XAI approaches since it may aid financial organizations in making effective judgments. In this study, we showed how to understand predictions from a random forest model using two well-known interpretable machine learning techniques, SHAP and LIME. Through SHAP, we were able to pinpoint the key elements that determine whether a loan is granted or denied. On the other hand, LIME gave us local justifications that made it clear to us why a certain situation was labeled as a loan accepted or not. Combining these methods allowed us to better understand the model's decision-making process, which can help the model perform better and increase user confidence decisions.In the end, using interpretable machine learning and XAI approaches in loan approval can result in fairer, more precise, and more transparent decision-making, which will be advantageous to both the lenders and the loan applicants.

References

[1] C. S. Reddy, A. S. Siddiq and N. Jayapandian, ”Machine Learning based Loan Eligibility Prediction using Random Forest Model,” 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1073-1079, doi: 10.1109/ICCES54183.2022.9835875.

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

[3] https://arxiv.org/abs/2110.10790

[4] R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, Y. Levron, Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities, Energy and AI, Volume 9,2022,

[5] Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. 2020, [Online]. Available: arXiv:2006.11371.

[6] P Ratadiya, K Asawa, O Nikhal - arXiv preprint arXiv:2011.10981, 2020 - arxiv.org

[7] L. Lai, ”Loan Default Prediction with Machine Learning Techniques,” 2020 International Conference on Computer Communication and Network Security (CCNS), Xi'an, China, 2020, pp. 5-9, doi: 10.1109/CCNS50731.2020.00009.

[8] C. Naveen Kumar, D. Keerthana, M. Kavitha and M. Kalyani, ”Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector,” 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1007-1012, doi: 10.1109/ICCES54183.2022.9835725.

[9] R. Karthiban, M. Ambika and K. E. Kannammal, ”A Review on Machine Learning Classification Technique for Bank Loan Approval,” 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2019, pp. 1-6, doi: 10.1109/ICCCI.2019.8822014.