MACHINE LEARNING EXPLAINABILITY IN LOAN TRANSPARENCY DETERMINATION

Emma Quyen Do
University of California, Berkeley
Berkeley, CA, USA
dohoangthucquyen2004@gmail.com

PhD. Olfa Nasraoui
University of Louisville
Louisville, KY, USA
olfa.nasraoui@gmail.com

Kyle Spurlock
University of Louisville
Louisville, KY, USA
kspurlock2000@gmail.com

Abstract

Machine Learning (ML) models have become increasingly influential, ranging from product recommendations to intricate medical diagnoses. However, their complexity and lack of transparency pose significant challenges, especially in terms of accountability and interpretability. To address these issues, Explainable Artificial Intelligence (XAI) emerged, aiming to make ML models more understandable and interpretable. This paper explores the application of LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) to a random forest classifier model for predicting loan defaults. Utilizing a dataset from Lending Club, our objective is to identify the crucial factors influencing a borrower's likelihood of defaulting on a loan. While our findings reveal that the random forest model achieves high accuracy, we observed a concerning reliance on geographical features, highlighting potential biases. This underscores the imperative necessity of XAI techniques in ensuring ethical and responsible use of AI in the financial sector. Future research will focus on enhancing model interpretability, validating our findings, and addressing the limitations of post-hoc explanation methods like LIME to minimize misinterpretations. By seamlessly integrating robust ML models with fundamental XAI principles, we can significantly improve the transparency, accountability, and trustworthiness of AI systems.

ACM Reference Format:
Emma Quyen Do, PhD. Olfa Nasraoui, and Kyle Spurlock. 2018. MACHINE LEARNING EXPLAINABILITY IN LOAN TRANSPARENCY DETERMINATION . In *Proceedings of (Research Experience for Undergraduate (2024 Louisville, KY)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Machine Learning (ML) models have become increasingly pervasive in our daily lives, influencing everything from product recommendations to medical diagnoses. As AI becomes more integrated into our society, it is crucial to consider the ethical implications and potential risks associated with its use. One of the most significant concerns about ML models is their lack of transparency. The complex algorithms and vast amounts of data used to train these models often make it difficult to understand how they arrive at their decisions. This lack of transparency can pose risks, as it can be challenging to hold AI systems accountable for their actions or to identify and correct any biases or errors in their decision-making processes. [2]

To address these concerns, Explainable Artificial Intelligence (XAI) aims to develop techniques that make ML models more interpretable and understandable. XAI approaches can provide insights into how models work, the factors that influence their decisions, and the level of uncertainty associated with their predictions. By making ML models more transparent, XAI can help to build trust in AI systems and ensure that they are used responsibly and ethically. [4] Several different approaches can be used to achieve XAI. One common approach is to use feature importance techniques to identify the input features that have the greatest impact on a model's predictions. Another approach is to use visualization techniques to represent the decision-making process of a model in a visually intuitive way. In this paper, we will be using LIME (Local-agnostic Model Explanator) as well as SHAP (SHapley Additive exPlanations) to explain the financial decision of deciding whether a user is going to be a loan default or not.

The development of XAI is an ongoing area of research, and there are still many challenges to overcome. However, the increasing importance of AI in our society makes it essential to develop techniques that can make AI systems more transparent, accountable, and trustworthy. By combining the power of AI with the principles of XAI, we can unlock the full potential of AI while also ensuring that it is used ethically and responsibly.

2 Local Agnostic Models

In the realm of Explainable Artificial Intelligence (XAI), local agnostic models hold a pivotal role by offering interpretable explanations for individual predictions made by complex machine learning models. Termed "agnostic," these models operate without requiring access to the inner workings or parameters of the predictive model, treating it as an opaque black box. This approach grants local agnostic models universal applicability across diverse machine learning models, encompassing deep neural networks and ensemble methods. Their primary aim is to shed light on the reasoning behind specific predictions, providing granular insights into how each feature contributes to a particular outcome. [1]

Among the notable techniques in this domain is LIME (Local Interpretable Model-agnostic Explanations). LIME generates local surrogate models by systematically perturbing input features and observing the resulting changes in predictions. This process enables LIME to infer the relative importance and influence of each feature on the model's decision for that specific instance. The localized

Research Experience for Undergraduate (2024 Louisville, KY,
,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

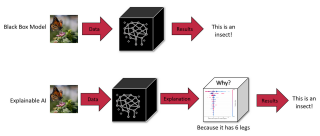

Figure 1: Explanability in AI framework. Source: https://encord.com/blog/model-robustness-machine-learning-strategies/

focus of local agnostic models empowers stakeholders to comprehend and have confidence in complex model predictions, bolstering the transparency and accountability of machine learning systems in critical applications.

2.1 Understanding LIME

Lime (Local Interpretable Model-agnostic Explanations) is a prominent technique designed to elucidate the predictions of complex machine learning models. It is inherently model-agnostic, meaning it can be applied to any machine learning model, regardless of its architecture or complexity. This flexibility makes it a valuable tool for understanding the inner workings of various machine learning models, enabling practitioners to gain insights into their predictions and improve their overall interpretability.

The core principle of LIME involves treating the model as a black box, requiring access solely to the input data and the model's output predictions. To generate local explanations, LIME employs a strategy of perturbing the input data around the instance being explained. This perturbation process involves creating a set of synthetic data points by making slight alterations to the features of the original instance. By carefully controlling the magnitude and direction of these perturbations, LIME aims to capture the local behavior of the model in the vicinity of the instance of interest.

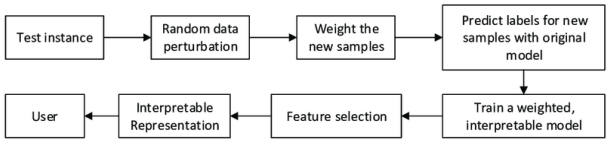

Figure 2: LIME framework within a black box model. (https://www.researchgate.net/figure/A-block-diagram-of-the-LIME-framework_fig1_352925192).

Once the synthetic data points are generated, LIME uses the original model to predict outcomes for these perturbed instances. These predictions, along with the perturbed data points, form a new dataset that LIME uses to fit an interpretable model, such as a linear model or decision tree. This interpretable model approximates the behavior of the complex model in the vicinity of the specific instance being analyzed. By leveraging simpler, more interpretable models, LIME can provide insights into the features and relationships that influence the predictions of the original model.

$$\xi(x) = \arg\min_{g \in G} \left(\mathcal{L}(f, g, \pi_x) + \Omega(g) \right)$$

where:

- $\xi(x)$ represents the explanation model for the instance x,
- G is the family of possible interpretable models (e.g., linear models),
- $\mathcal{L}(f, g, \pi_x)$ is a measure of how unfaithful g is in approximating f in the locality defined by π_x,
- π_x is a proximity measure that defines the locality around x,
- $\Omega(g)$ is a complexity measure of the interpretable model g.

To ensure the explanations are locally faithful, LIME assigns weights to the perturbed instances based on their similarity to the original instance. These weights are typically determined using a distance metric, such as the Euclidean distance, ensuring that data points closer to the original instance have a greater influence on the explanation. This weighting scheme helps LIME focus on the most relevant features for the prediction at hand, providing more accurate and reliable explanations.

LIME has been instrumental in various applications, from healthcare to finance, where understanding model predictions is crucial. In healthcare, LIME has been used to explain the predictions of machine learning models used for disease diagnosis, treatment planning, and drug discovery. By providing interpretable explanations, LIME helps medical professionals trust and verify the decisions made by these models, ultimately leading to improved patient care and outcomes. In finance, LIME has been used to explain the predictions of models used for credit scoring, fraud detection, and investment recommendations. By understanding the factors that influence these predictions, financial institutions can make more informed decisions, mitigate risks, and comply with regulatory requirements. [1]

The ability to generate transparent and comprehensible explanations without compromising the performance of the original model makes LIME a valuable tool in the pursuit of Explainable Artificial Intelligence (XAI). By empowering stakeholders to understand how and why machine learning models make predictions, LIME promotes trust, accountability, and ethical use of AI systems in various domains.

2.2 Understanding SHAP

SHAP (SHapley Additive exPlanations) serves as a powerful tool for deciphering the intricate workings of machine learning models, providing valuable insights into the model's behavior and the significance of individual features. At its core, SHAP is rooted in cooperative game theory, which centers around the Shapley value concept. This concept ensures that feature contributions are fairly distributed, considering their impact across all possible feature combinations. This approach ensures that the explanations generated by SHAP are not only locally accurate but also globally consistent, leading to reliable and comprehensive interpretations.

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(x_S \cup \{i\}) - f(x_S) \right]$$

where:

- $\phi_i(f, x)$ is the SHAP value for feature i,
- N is the set of all features,
- S is a subset of features not including i,

- $|S|$ is the number of elements in subset S,
- $f(x_S \cup \{i\})$ is the model output with feature i included in subset S,
- $f(x_S)$ is the model output with only the features in subset S.

One of the key advantages of SHAP is its ability to handle complex, non-linear models. In such scenarios, traditional interpretability methods often fall short, as they struggle to capture the intricate relationships between features and model outputs. SHAP, on the other hand, excels in these situations, effectively providing clear and intuitive explanations. This is because SHAP calculates the contribution of each feature by considering its impact on the model's output, taking into account interactions and dependencies among features.

Moreover, SHAP offers a holistic view of the model's behavior by aggregating the contributions of individual features. This aggregation process allows researchers and practitioners to understand how different features collectively influence the model's predictions. This comprehensive analysis enables the identification of key drivers behind the model's decisions, helping to uncover patterns and relationships that might not be immediately apparent from examining individual features in isolation.

The local accuracy of SHAP ensures that explanations are tailored to specific instances or predictions. This means that the contributions of features are evaluated for each data point or observation, providing nuanced insights into how the model arrived at its conclusions. This local focus allows for a deeper understanding of the model's behavior across different input values and scenarios.

In summary, SHAP serves as an invaluable tool for interpreting the output of machine learning models by quantifying the contribution of each feature and providing clear and intuitive explanations. Its robustness, ability to handle complex models, and emphasis on both local and global consistency make it a trusted resource for researchers and practitioners seeking to gain insights into the underlying mechanisms driving model predictions.

3 Experiments

In this experiment, we will be using a dataset obtained from Kaggle.com, named "Lending Club Loan Defaulter." Our primary objective is to investigate whether we can discern the factors that influence a borrower's likelihood of becoming a loan defaulter or, conversely, their propensity to repay the loan in full.

These two possible outcomes are the class labels to be predicted by the ML model, and the frequency of each class label is depicted in Figure **??**

We will use a black box model called Random Forest. This black-box model, available in the sci-kit learn library, is renowned for its ability to handle complex, high-dimensional datasets and uncover intricate patterns and relationships within them.The Random Forest model operates on the principle of ensemble learning, which combines the predictions of multiple decision trees to arrive at a final, more accurate prediction. Each decision tree is trained on a subset of the data and makes its own prediction. The final prediction is determined by aggregating the predictions of all the decision trees, typically by majority vote.[3]

We aim to uncover the key factors that contribute to loan default by implementing LIME to determine the weight of each feature to the final decision. These factors may include variables such as the borrower's credit history, income, debt-to-income ratio, loan amount, and loan term, among others.

Once the Random Forest model has been trained on a portion of the dataset, we will evaluate its performance on a separate holdout set. This process, known as cross-validation, helps us assess the model's ability to generalize to unseen data and provides a more reliable estimate of its predictive accuracy.

Through this experiment, we hope to gain valuable insights into the factors that influence loan default and demonstrate the effectiveness of Random Forest in addressing this challenging problem. By understanding the underlying patterns and relationships, lenders can make more informed decisions, mitigate risk, and improve their lending practices.

4 Handling the Data

The dataset we will be utilizing is composed of 396030 data points, each characterized by 27 unique features. These features encompass a wide range of information relevant to loan applications, including the loan amount, interest rate, annual income, debt-to-income ratio, and the number of open accounts.

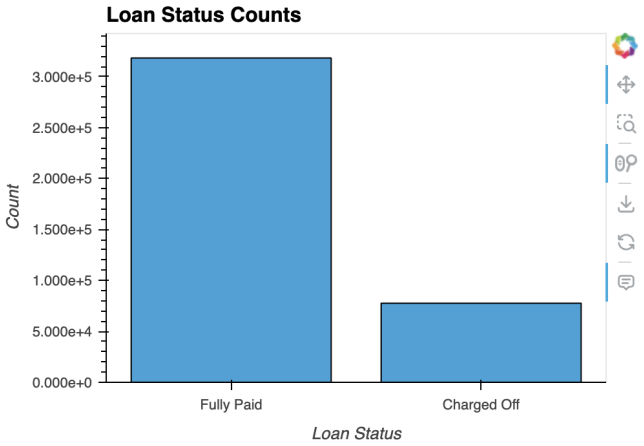

Figure 3: The count of fully paid and charged off loan borrowers.

It is crucial that we preprocess the data to ensure its accuracy and completeness. One of the key steps in this process involves decoding the address information into three distinct components: zip region, zip code, and zip local. This step allows us to extract valuable insights from the address data, such as the geographic location of the loan applicants.Additionally, we dropped columns that have an excessive number of null values. Null values can introduce bias and uncertainty into our models, so it is essential to remove them to maintain data integrity.

We also encoded the loan sub-grade value for each data point. The loan sub-grade value is a categorical variable that indicates the creditworthiness of the loan applicant. By encoding this variable, we can transform it into a numerical format that can be easily processed by our models.

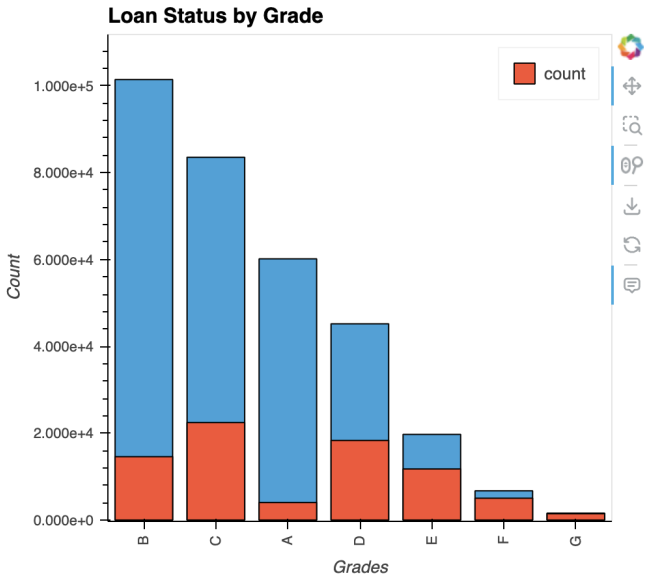

Figure 4: The comparison between loan borrowers with different grade.

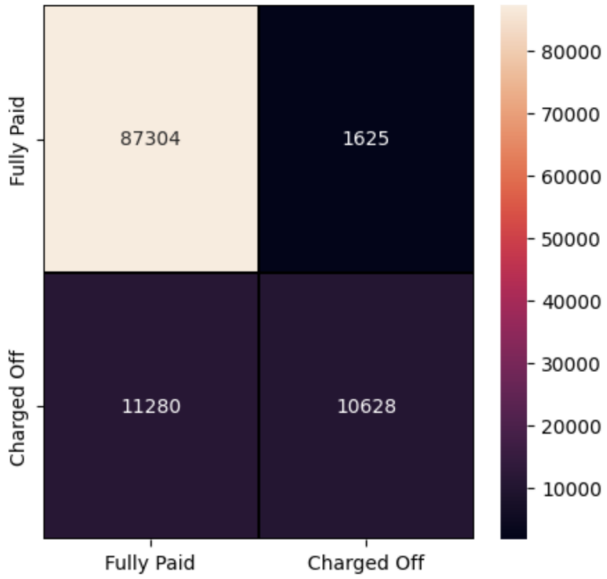

Figure 5: The confusion matrix for the dataset

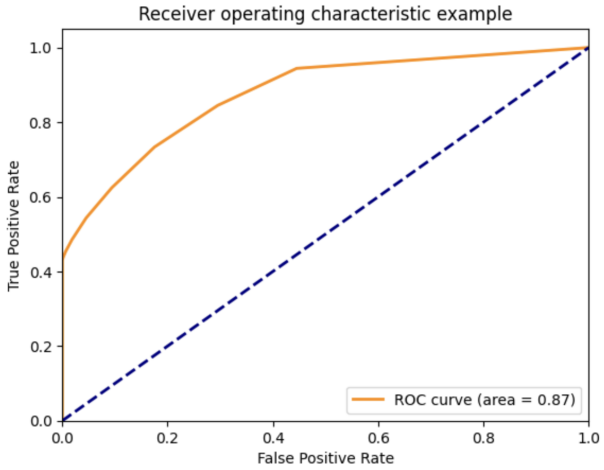

Figure 6: The ROC of the data set

Once the data preprocessing is complete, we proceeded to split it into training and testing sets. The test set comprised 33% of the overall data, while the remaining 67% will be allocated to the training set. This split ensures that we have a representative sample of the data for both training and testing purposes.

Finally, we generated a unique random state with the Random Forest black box model. The random state is a crucial parameter that controls the randomness of the model. By generating a unique random state, we can ensure that our model is not influenced by previous runs and that the results are reproducible.

The Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) of a classification model at various classification thresholds. The TPR, also known as sensitivity or recall, measures the proportion of actual positive cases that are correctly identified by the model, while the FPR, also known as the false alarm rate, measures the proportion of actual negative cases that are incorrectly classified as positive.

The ROC curve is constructed by plotting the TPR against the FPR for different classification thresholds. The threshold determines the decision boundary for classifying an instance as positive or negative. As the threshold increases, the TPR typically decreases while the FPR increases. This is because a higher threshold makes it more difficult for positive instances to be classified correctly while making it easier for negative instances to be classified correctly.

The area under the ROC curve (AUC) provides a measure of the overall performance of a classification model. An AUC of 1 indicates perfect classification, where all positive instances are correctly identified and all negative instances are correctly rejected. An AUC of 0.5 indicates random chance, where the model is no better than flipping a coin to make a classification.

In the given context, the ROC analysis suggests that the model under consideration performs better than random chance and reaches higher accuracy when the false positive rate is relatively small compared to the true positive rate. This indicates that the model is able to correctly identify a significant proportion of positive instances while maintaining a low rate of false positives. This is a desirable characteristic for classification models, particularly in applications where false positives can have significant consequences.

5 Interpreting the Result

After training our model, we implemented LIME (Local Interpretable Model-agnostic Explanations) on our training data to gain insights into the features influencing our model's decisions. We selected a

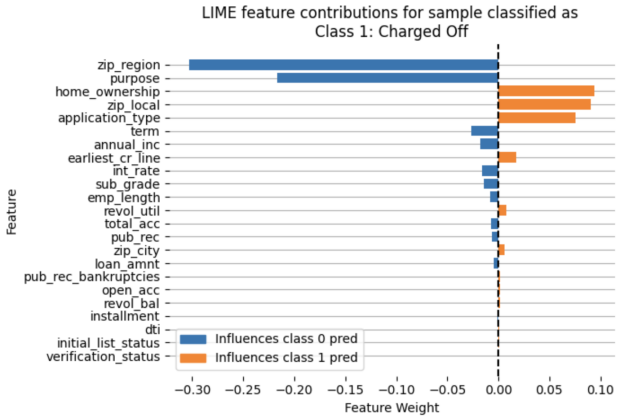

Figure 7: The weight of features in the charged off model

data point classified as "charged off" to understand the factors contributing to this outcome. The analysis revealed that the primary determinants for loan repayment decisions were predominantly based on the borrower's zip code, city, and state. Our black box model yielded a high accuracy rate of 98%. However, this result underscores the critical importance of Explainable AI (XAI). Without employing a local model-agnostic approach like LIME to interpret our model's decisions, we would not have uncovered the significant bias embedded within the model. The model's reliance on geographical factors such as zip code and state indicates a potentially unfair and unethical bias in determining loan eligibility. More specifically, the fact that certain applicants' living address (as inferred from their zip code) plays a crucial factor in rejecting (or accepting) their loan is a form of financial redlining (or privilege in the opposite/acceptance case). This discovery raises serious ethical and fairness concerns in the context of lending decisions. It highlights the necessity for transparency and accountability in AI models, particularly in finance, where decisions have substantial real-world impacts. Ensuring that models do not perpetuate or exacerbate existing biases is crucial for maintaining fairness and trust in automated decision-making systems.

6 Conclusion

The outcome of our loan default data, which was trained using a black box model, raises significant ethical concerns. While the model may exhibit high accuracy in predicting loan defaults, the features it relies upon for making these predictions are deeply problematic. These features introduce biases that have no legitimate bearing on an individual's creditworthiness or ability to repay a loan.

The current model faces a significant limitation due to its post-hoc nature. LIME (Local Interpretable Model-agnostic Explanations) generates proxy explanation models to approximate the behavior of a previously trained black box model using data that wasn't part of the original analysis. This approach may not fully capture the underlying relationships in the original model, leading to misinterpretations. Additionally, LIME explanations can be complex and

Sample: 368351, Actual: 1, Pred: 1

zip_region	9
zip_local	00
home_ownership	RENT
term	36
int_rate	9.76
sub_grade	7
annual_inc	75000.0
zip_city	37
revol_util	85.3
loan_amnt	8000.0
verification_status	Source Verified
earliest_cr_line	2006
emp_length	9.0
pub_rec	0.0
purpose	debt_consolidation
dti	19.42
installment	257.24
open_acc	10.0
revol_bal	20802.0
pub_rec_bankruptcies	0.0
total_acc	21.0
application_type	INDIVIDUAL
initial_list_status	w

Figure 8: The count of fully paid and charged off loan borrowers.

challenging to interpret, increasing the risk of overinterpretation and incorrect conclusions.

Future research should focus on enhancing the model by training it on advanced black box models like XGBoost to improve the accuracy and robustness of the explanations. It's also crucial to calculate fairness metrics such as disparate impact, equal opportunity difference, and demographic parity to ensure that the model's predictions don't discriminate against specific groups, making it fairer and ethically sound.

Developing new methods for explaining black box models that are more accurate, robust, and easier to interpret will enable users to better understand these models and make more informed decisions about their use. Exploring the use of LIME for other model types like deep learning models will expand its versatility in explaining a broader range of complex machine learning models.

Acknowledgments

This research paper and project was developed during the Research for Undergraduate Experience conducted at University of Louisville, Kentucky during May-July of 2024. This research was supported by the U.S. National Science Foundation (NSF) under grant CNS-2050925.

References

[1] Tsehay Admassu Assegie. 2023. Evaluation of local interpretable model-agnostic explanation and shapley additive explanation for chronic heart disease detection. *Proc Eng Technol Innov* 23 (2023), 48–59.

[2] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[3] Lilian Weng. 2017. How to Explain the Prediction of a Machine Learning Model? *Blog Post, https://lilianweng. github. io/lil-log/2017/08/01/how-to-explain-the-prediction-of-a-machine-learning-model. html* (2017), 1–7.

[4] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8*. Springer, 563–574.