Researchers: Shijie Han, David Jay Maltz, Rui Tang
Mentors: Dr. Ali Hirsa, Dr. Branka Hadji, Miao Wang

Explainable AI: Improving the Stability of LIME

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Abstract

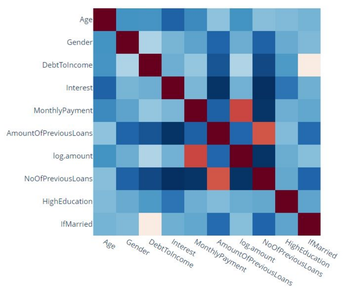
This study advances credit risk assessment in Peer-to-Peer (P2P) lending by integrating traditional financial data with machine learning and explainable AI techniques. We optimize prediction performance using advanced machine learning techniques like XGBoost and GLM while enhancing transparency through SHAP and LIME. Focusing on LIME's limitations, such as instability and sensitivity to hyperparameters, we implement and evaluate enhanced versions including S-LIME, DLIME, OptiLIME, and BayLIME. These improvements aim to increase the stability and reliability of explanations in P2P lending risk assessment.

Introduction

The rise of Peer-to-Peer (P2P) lending has transformed financial services, but its unique risks necessitate robust risk management. Machine learning (ML) has revolutionized risk management, yet the black-box nature of ML models poses challenges, particularly in the regulated financial industry. Explainable AI (XAI) is crucial to foster trust, ethics, and compliance. This work focuses on improving the LIME technique, a prominent XAI method, to enhance its stability and reliability for practical application in P2P lending risk management.

Preliminary Works

The study used the Bondora P2P lending dataset, which includes over 30,000 records with 190 features and a binary default label. The data was carefully preprocessed, handling missing values, detecting and treating outliers, scaling features, and encoding categorical variables. Exploratory data analysis identified 10 key predictive features, including age, gender, debt-to-income ratio, interest rate, monthly payment, and loan history. Descriptive statistics revealed diverse borrower and loan characteristics, with notable correlations between the features. An XGBoost model was then trained using the selected features, and feature importance analysis highlighted the critical role of financial variables like loan amount, monthly payment, and previous loans, as well as demographic factors like age and education level. To further explain the model's predictions, the researchers conducted SHAP and LIME analyses, aiming to enhance the interpretability and transparency of the credit risk assessment framework for P2P lending.



Introduction on LIME

The use of XAI modules is essential for elucidating the decision-making process of black-box machine learning models, particularly in the context of loan issuance. These modules offer both stakeholders a clear and impartial explanation by analyzing the relative importance of the various factors influencing the loan decision. This need for transparency is well-addressed by the LIME model, which explores the local behavior of an instance by generating perturbed samples in its vicinity and fitting an interpretable model to predict the instance's decision. Each perturbation is evaluated based on its proximity to the original instance and the complexity of the interpretable model. The local model can be from the class of potentially interpretable models such as linear models, decision trees, etc. The explanations provided by LIME for each observation x is obtained as follows:

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

where G is the class of potentially interpretable models such as linear models and decision trees

The goal is to minimize the locality aware loss L without making any assumptions about f , since a key property of LIME is that it is model agnostic. L is the measure of how unfaithful g is in approximating f in the locality defined by $\pi(x)$.

Proportionality of LIME Stability and Model Complexity

While LIME offers an understandable and unbiased explanation of black-box ML models, its reliance on generating random perturbations around an instance raises concerns about its stability. This instability becomes particularly evident when comparing interpretable models of different complexities. To assess this, four models with varying complexities were analyzed to determine feature rank instability, feature value instability, and Jaccard index rankings for the top five features using the P2P Loan dataset. The rank and value instability were measured by calculating the variance of each feature's ranking and value across 20 iterations. The models, listed from lowest to highest complexity, included Logistic Regression, Decision Tree, Random Forest, and Neural Network. In below tables, notice the relative increase in instability for both the features' rank and value as the model complexity increases. There seems to be less of a conclusive pattern with the Jaccard index with respect to instability with complexity especially past the second index; however, it is noticeable that generally as model complexity increases, the Jaccard ranking decreases.

Index	Logistic Regression	Decision Tree (depth=3)	Random Forest (n_estimators=10)	Neural Network (1 layer, 10 units)
1	0.89578947	0.92578947	0.54131579	0.66789474
2	0.78394737	0.92666667	0.5227193	0.72877193
3	0.72263158	1.	0.66307895	0.749
4	0.74568421	0.70778947	0.77318296	0.73950877
5	0.79626566	0.60037594	0.74156955	0.74232038

	Logistic Regression	Decision Tree (depth=3)	Random Forest (n_estimators=10)	Neural Network (1 layer, 10 units)
Rank Instability	0.118	0.959	1.629	1.821
Value Instability	2.155e-05	7.164e-05	0.00269	0.00544

Examining Robustness with LIME Stability

The paper "On the Robustness of Interpretability Methods" proposes a framework to assess the robustness of interpretability methods, such as LIME, by evaluating how small input modifications affect the resulting explanations. The key steps are:

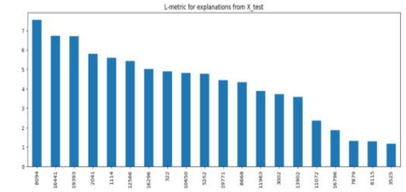
1. Feed the machine learning model and the interpretability method (e.g., LIME) to obtain feature contributions.
2. Introduce small, random perturbations to the input and observe the changes in the explanations.
3. Calculate a Lipschitz value that quantifies the sensitivity of the explanations to input changes.

The Lipschitz value is defined by the equation:

$$\tilde{L}_X(x_i) = \arg \max_{x_j \in \mathcal{N}_\epsilon(x_i) \leq \epsilon} \frac{\|\|f(x_i) - f(x_j)\|\|_2}{\|\|x_i - x_j\|\|_2}$$

where

$$\mathcal{N}_\epsilon(x_i) = \{x_j \in X \mid \|\|x_i - x_j\|\|_2 \leq \epsilon\}$$



By comparing the average Lipschitz values across different interpretability methods, or by ranking instances based on these values, one can identify the most robust or unstable explanations within a dataset (see above figures for robustness results using P2P Loan dataset).

Enhancing Stability in LIME

To address the instability of LIME, several methods have been proposed:

1. Stabilized-LIME: This approach leverages the Central Limit Theorem to automatically determine the optimal number of perturbations needed to ensure the stability of LIME explanations for LARS/LASSO regression models. It incorporates hypothesis testing to verify if the number of perturbations is sufficient.
2. DLIME: A deterministic version of LIME that replaces random perturbations with Agglomerative Hierarchical Clustering and K-Nearest Neighbor to select relevant training data points. This ensures consistent and stable explanations, although the quality depends on the clustering process.
3. OptiLIME: This method addresses LIME's instability through two key improvements:
   a. Eliminating the ridge penalty, as LIME's data generation process inherently produces points on the model's surface, making the ridge penalty unnecessary.
   b. Optimizing the kernel size used in the locality of the LIME explanation to balance the trade-off between stability (measured by CSI/VSI scores) and adherence (measured by R^2 score).
4. BayesLIME: Presented in the paper "Reliable Post hoc Explanations: Modeling Uncertainty in Explainability," this framework offers a more compelling Bayesian approach. BayesLIME provides a way to quantify the uncertainty in LIME's feature importance scores by modeling them as probability distributions and generating confidence intervals. This allows for a more reliable interpretation of the explanations, accounting for the inherent uncertainty in the LIME process.

These approaches aim to improve the reliability and consistency of LIME explanations, which is crucial for their practical application, especially in sensitive domains like finance.

Conclusions

In light of the increasing reliance on machine learning models for critical lending decisions in P2P lending, it is crucial for all parties involved to understand the reasoning behind these model predictions. Explainable AI models, such as SHAP and LIME, offer frameworks to assess and rank the importance of features relevant to loan assessments. LIME, in particular, generates local perturbations around an instance and fits an interpretable model to determine feature weightings. However, this approach raises concerns regarding the stability and robustness of the results, especially as instances or the complexity of the interpretable model vary. To address these concerns, several enhanced versions of LIME have been proposed, including S-LIME, DLIME, OptiLIME, and BayLIME. These models aim to improve the stability and reliability of LIME, offering users more dependable insights into model decisions. By advancing these methods, we move closer to achieving more trustworthy and consistent explanations, which are essential for informed decision-making in the lending industry.