|  |
| --- |
| **BRIDGING GAPS IN EXPLAINABLE AI FOR FINANCE: MULTI-STAKEHOLDER ANALYSIS** |
|  |
| **Shreya Palande, Parkash Singh** |

Introduction

The rapid adoption of AI models in the financial industry has raised critical concerns about their transparency, trustworthiness, and interpretability. These concerns are particularly significant as decisions made by these models directly affect stakeholders such as customers, loan officers, and auditors. Explainable AI (XAI) techniques aim to address these issues by providing actionable insights into model predictions. However, existing methods often fail to cater to the specific needs of diverse stakeholders, making it challenging to interpret and trust the results.

This research focuses on bridging the gaps in explainable AI by leveraging advanced techniques like C-CHVAE, LIME, SHAP, and Protodash. Using the HELOC dataset, this project aims to analyze and compare these methods in providing tailored explanations for different stakeholders. The primary goal is to achieve state-of-the-art performance on the Home Equity Line of Credit (HELOC) dataset while improving transparency, fostering trust, and ensuring regulatory compliance in financial decision-making processes.

Related Work

Explainable AI has gained significant attention across various domains, including medicine, transportation, and finance. The methods vary based on their scope (local vs. global), type (model-specific vs. model-agnostic), and application area. Existing XAI techniques primarily fall into three categories:

1. Feature Relevance Explanations: Techniques like SHAP and LIME provide insights into the importance of features influencing individual predictions.
2. Simplified Models: Approaches using global interpretable models for auditing and compliance.
3. Counterfactual Explanations: Techniques like C-CHVAE generate actionable recommendations by identifying feature changes required to achieve the desired outcome.

In the financial domain, most studies focus on local explanations, as they allow stakeholders to understand individual decisions. However, challenges remain in adapting these methods to meet the needs of different stakeholders, evaluating the quality of explanations, and addressing data security concerns.

The study by Demajo et al. (2020) introduces an explanation framework for AI-driven credit scoring that incorporates global, local feature-based, and local instance-based explanations. Using the XGBoost algorithm, they achieved state-of-the-art performance on the HELOC dataset, highlighting the importance of transparency for various stakeholders. This aligns closely with our research, which employs a similar multi-faceted approach to address the diverse requirements of stakeholders in financial decision-making.

As shown in this recent paper published in July 2024 (\*Explainable Artificial Intelligence (XAI) in Finance: A Systematic Literature Review\* by Jurgita Černevičienė and Audrius Kabašinskas), the authors identified three main challenges in conducting comprehensive reviews. These include difficulties in determining the intended audience for the explanation (technical specialists vs. end users), the lack of new XAI techniques or explainable machine learning methods to tackle financial problems, and the absence of explainability valuation metrics, which play a significant role in building trust in AI models. Furthermore, ensuring the security of information provided to the XAI model remains a critical concern.

This study aims to fill these gaps by targeting techniques for different stakeholders involved in credit scoring. Although a universal evaluation matrix for XAI techniques is still lacking, our focus is on how these methods assist decision-making in finance. Interpretability should be judged by the practical utility of these techniques, and future studies will extend this usability analysis to other fields like healthcare and transportation.

Methodology

This research applies a multi-stakeholder approach to explainable AI (XAI) to enhance the transparency, fairness, and interpretability of machine learning models used in credit risk prediction. The study focuses on the HELOC dataset, which contains customer information used to predict credit risk as "Good" or "Bad". The methodology comprises data preprocessing, model training, the application of various XAI techniques, and evaluation based on specific stakeholder needs. Below is a detailed description of the methodology employed in the study.

A diagram of a diagram

Description automatically generated

The above flowchart depicts the proposed methodology.

1. **Data Preprocessing**

This project uses the HELOC dataset which contains anonymized credit data used to assess credit risk, classifying applicants as "Good" or "Bad" based on repayment behavior. It includes features like trade counts, credit utilization, and payment history, providing insights into financial reliability. The dataset is widely used for building and evaluating predictive financial models.

A screenshot of a computer

Description automatically generated

The preprocessing phase ensures that the data is ready for model training by handling missing values, normalizing features, and splitting the data for training and testing. The following steps were taken:

* Handling Missing Values: Missing or special values (such as -9, -8, and -7) in the dataset were identified and replaced with NaN to enable proper handling.
* Imputation: Numerical columns with missing values were imputed using the SimpleImputer method with a mean imputation strategy, replacing missing values with the column mean.
* Feature Normalization: Features were normalized using StandardScaler to ensure that all features are on a comparable scale, as this helps improve model performance and interpretability.
* Splitting the Data: The data was split into training and testing sets using a 80-20 split, ensuring that the training set was used to fit the models and the testing set was reserved for evaluation.

The dataset was processed into X (features) and y (target variable, "RiskPerformance") for further model training.

1. **Model Training**

Two powerful gradient boosting models, XGBoost and LightGBM, were trained on the preprocessed dataset. These models were chosen for their performance with structured/tabular data:

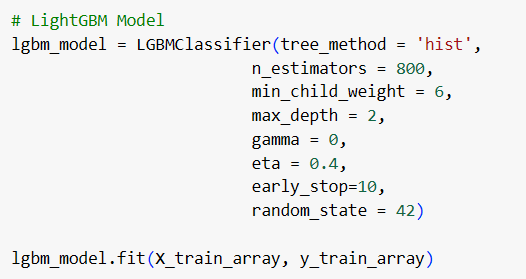
* XGBoost: This gradient boosting algorithm was trained using the XGBClassifier, which learns a set of decision trees in a sequential manner, where each tree tries to correct the errors of the previous one.

The following parameters were set for XGBoost algorithm:

A screenshot of a computer code

Description automatically generated

* LightGBM: Similar to XGBoost, LGBMClassifier was used for training. LightGBM is designed to handle large datasets efficiently and often performs better than other models in terms of speed and accuracy.



The models were trained on the training data (X\_train) and evaluated on the test data (X\_test) using accuracy and other performance metrics.

1. **Explainability Techniques**

To provide actionable insights for stakeholders, a variety of XAI techniques were applied to explain the predictions of the trained models. The following techniques were used:

* C-CHVAE (Counterfactual Variational Autoencoder):

C-CHVAE was used to generate counterfactual explanations for individual customers. This technique identifies the minimal changes needed to alter a customer's credit risk prediction. These changes, such as reducing credit utilization or increasing income, provide actionable recommendations to customers for improving their creditworthiness. Counterfactuals were generated for customers labeled as “Bad” to suggest feasible actions for improving their credit risk to “Good.”

* LIME (Local Interpretable Model-agnostic Explanations):

LIME was applied to generate local explanations for individual predictions. This method approximates the model's behavior locally by fitting a simple interpretable model (like a linear model) to the dataset around a specific instance.

The goal was to help stakeholders understand the factors influencing a particular customer's prediction, offering them explanations about why a decision was made.

* SHAP (SHapley Additive exPlanations):

SHAP was used to provide both local and global explanations. Local explanations indicate how individual features contributed to a specific prediction, while global explanations summarize feature importance across the entire dataset.

SHAP values were computed to visualize which features (e.g., credit utilization, income) played the most significant roles in the credit risk prediction for all instances.

* ProtoDash:

ProtoDash was used to generate prototype-based explanations, which simplify the decision-making process for loan officers. This technique compares a given instance to representative prototypes of Good and Bad outcomes, making it easier for loan officers to justify decisions based on the closeness to these prototypes

Stakeholder-Specific Focus

The research identifies and addresses the needs of different stakeholders by tailoring the explanations to their specific roles:

* Customers: C-CHVAE provides customers with actionable counterfactuals—practical suggestions for improving their credit risk (e.g., "increase your income by $500 to qualify for a better loan offer").
* Loan Officers: ProtoDash provides prototype-based explanations that help loan officers justify decisions in an understandable and transparent manner. By comparing a customer's profile to typical "Good" and "Bad" cases, loan officers can rationalize their decisions in credit risk assessments.
* Technical Teams: SHAP and LIME provides detailed insights into feature relevance, which is useful for technical teams who need to understand the model’s behavior and ensure its fairness and transparency.

Evaluation

**Model Performance:**

XGBoost and LightGBM performance metrics (accuracy, F1-score, etc.) are evaluated:

**A screenshot of a test

Description automatically generatedXGBoost**: **LightGBM**:   
A screenshot of a computer program

Description automatically generated

**Evaluation** **Metrics for XAI Techniques:**

1. Fidelity: Measure how accurately the explanations reflect the model’s behavior.

2. Actionability: Assess the practicality of recommendations for stakeholders, such as customers.

3. Realism: Evaluate the plausibility of counterfactual examples generated by C-CHVAE.

4. Transparency: Determine how well techniques like SHAP enhance the interpretability of model predictions.

5. Stakeholder Satisfaction: Collecting feedback on the usability of explanations for each stakeholder group will be part of future work.

Results

LIME

A screenshot of a computer

Description automatically generated

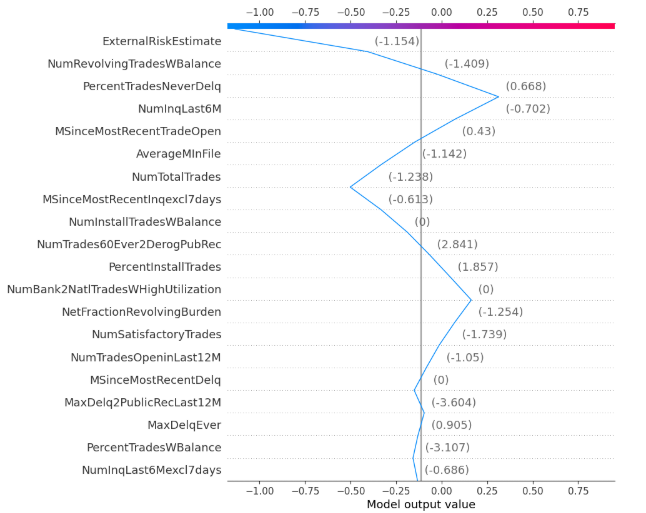
A screenshot of a computer screen

Description automatically generated

The LIME graph visually explains the model’s prediction for an individual instance (e.g., credit risk prediction). The **Good** class (0.23) and **Bad** class (0.77) show the predicted probabilities. Features are ranked by their influence, with **NetFractionRevolvingBurden** and **PercentTradesNeverDelq** being crucial for "Good" or "Bad" predictions.

For the **technical team**, this explanation reveals how features like **NumInqLast6M** or **NumSatisfactoryTrades** impact the model’s decision. They can identify critical features affecting predictions, ensuring the model behaves as expected and making it easier to improve or debug the model for fairness and transparency.

* 1. SHAP



The SHAP (SHapley Additive exPlanations) graph shows the contribution of each feature to the model’s output for a given instance, allowing for both global and local interpretability. The model output value is shown on the x-axis, with negative values indicating a prediction of "Bad" and positive values indicating "Good."

Key observations in this SHAP plot:

1. ExternalRiskEstimate significantly contributes to a negative output (towards "Bad"), as indicated by the blue color. Its large negative effect (-1.154) suggests a high risk.
2. Features such as NumRevolvingTradesWBalance, PercentTradesNeverDelq, and NumInqLast6M have varying degrees of influence, with PercentTradesNeverDelq contributing positively (towards "Good"), supporting a better credit profile.
3. The NumInstallTradesWBalance and NumTrades60Ever2DerogPubRec features are influential but have less impact compared to others.

For the technical team, this SHAP plot helps understand which features are driving the prediction (whether for "Good" or "Bad") and provides a transparent explanation for model output. They can focus on adjusting these key features for model improvement or debugging.

* 1. C-CHVAE

A screenshot of a computer

Description automatically generated

The C-CHVAE output provides a counterfactual example, showing the minimal changes needed in customer features to shift their credit classification from "Bad" to "Good." This helps customers understand actionable steps to improve their creditworthiness, aids loan officers in advising clients, and ensures model transparency for technical teams.

A graph of different colored rectangles

Description automatically generated

The heatmap generated for C-CHVAE compares the original and counterfactual instances of a customer's credit data. It highlights which features required the most significant changes to shift the classification from "Bad" to "Good." The colors represent the magnitude of the difference in feature values, with blue indicating a reduction and orange indicating an increase in value.

A black text on a white background

Description automatically generated

A number and numbers on a white background

Description automatically generated

C-CHVAE also generates automated output suggesting actionable changes to improve creditworthiness.

* 1. PROTODASH

A graph with different colored lines

Description automatically generated with medium confidence

The Protodash output provides a set of representative prototypes (or "nearest neighbors") that explain an individual's classification. Each prototype is a point in the feature space that is most similar to the instance being analyzed, showing a decision boundary. These prototypes act as human-readable justifications for the model's prediction, helping to identify the key characteristics influencing a customer's credit risk classification (e.g., "Good" or "Bad"). The heatmap generated by Protodash visualizes these prototypes against the input features, making it easier for loan officers to understand what aspects of a customer's credit behavior lead to their decision.

1. Conclusion

This research demonstrates the importance of tailoring XAI techniques to meet the specific needs of diverse stakeholders in the financial domain. By applying methods such as C-CHVAE, LIME, SHAP, and Protodash to the HELOC dataset, we address challenges related to transparency, trust, and usability in AI-driven decision-making.

1. Key Contributions

Shreya:

 Data Engineering & Processing

* Led feature engineering and preprocessing of the HELOC dataset
* Developed automated pipeline integrating XGBoost and LightGBM models with XAI techniques
* Enhanced and refined XAI implementation for improved performance

 Visualization & Documentation

* Created innovative visualization methods tailored for non-technical stakeholders
* Designed and created project poster with content input from Parkash
* Provided formatting and refinements to proposal and report

 Performance Optimization

* Benchmarked model performance against XAI explanation quality
* Analyzed accuracy-explanation tradeoffs
* Created automated explanation generation pipeline for enhanced scalability

Parkash:

 Model Development & Implementation

* Implemented and trained XGBoost and LightGBM models
* Developed core XAI technique implementations
* Conducted comprehensive analysis of XAI techniques for financial applications

 Research & Analysis

* Conducted extensive literature review identifying research gaps
* Formulated key research questions driving the project
* Developed stakeholder-specific categorization framework for explanations

 Documentation & Communication

* Authored primary content for project proposal and report
* Provided technical content for project poster
* Created recommendations for improving explainability metrics including fidelity and actionability

1. Future work

Future work will focus on:

1. Scaling evaluations to larger datasets.
2. Refining evaluation metrics for better assessment.
3. Addressing data security challenges in XAI for financial applications.
4. Utilizing NLP to process C-CHVAE for better assisting customer with counterfactuals.

References

1. Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable AI for Interpretable Credit Scoring. arXiv preprint arXiv:2012.03749. Retrieved from https://arxiv.org/abs/2012.03749

2. Černevičienė, J., & Kabašinskas, A. (2024). Explainable Artificial Intelligence (XAI) in Finance: A Systematic Literature Review. Retrieved from https://link.springer.com/article/10.1007/s10462-024-10854-8

3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Retrieved from https://arxiv.org/abs/1602.04938

4. Pawelczyk, M., Broelemann, K., & Kasneci, G. (2020). Counterfactual Explanations for Machine Learning on Tabular Data: A Survey. Retrieved from https://arxiv.org/abs/2001.07417