POWER OF EXPLAINABILITY

Alessandro Carella
MSc Student in Data Science & Business Informatics
University of Pisa
a.carella3@studenti.unipi.it

Rafael Ignacio Urbina Hincapie
MSc Student in Data Science & Business Informatics
University of Pisa
r.urbinahincapie@studenti.unipi.it

Sara Hoxha
MSc Student in Data Science & Business Informatics
University of Pisa
s.hoxha5@studenti.unipi.it

July 6, 2024

ABSTRACT

In this report, we aim to explore how Explainable AI enables the comprehension of black-box models and augments our human understanding of AI decisions. We implemented a black-box model and applied various XAI explainers to it, by focusing on evaluating across three dimensions: content, presentation, and user experience. Finally, we address fundamental research questions and draw conclusions about the effectiveness and impact of XAI.

1 Introduction

Nowadays, AI has become an integral part of our daily lives, influencing decision making in various domains. Yet, the lack of transparency in a majority of AI models' results poses a significant challenge in comprehending, and subsequently trusting their decisions. Explainable AI (XAI) aims to mend these issues by providing insights into how these models make their decisions, and this is precisely what we aim to do in this report. Specifically, we have selected and prepared a dataset, trained a black-box model, and employed various XAI methods to uncover how our model arrives at its decisions. The methodologies and evaluations of these explanations are discussed in detail in the following sections.

2 Data Understanding & Preparation

The dataset chosen for our study focuses on uplift modeling, which aims to identify customers most likely to respond positively to a marketing promotion. To be able to identify as such, we focused on two primary variables of interest: offer and conversion. The 'offer' variable represents the type of marketing promotion sent to customers, whereas the 'conversion' variable indicates whether the customer made a purchase.

Based on their values following the decision tree in Figure 1, we were able to define customer types which are usually four in uplift modeling:

1. Sure Things: Customers who would buy regardless of receiving a marketing promotion.
2. Persuadables: Customers who are more likely to buy the product when receiving a marketing promotion.
3. Sleeping Dogs: Customers who are less likely to buy when receiving a marketing promotion.
4. Lost Causes: Customers who would not buy regardless of receiving a marketing promotion.

As such, we can determine that it's beneficial to send marketing promotions only to Persuadables, as they're the sole segment where treating (sending promotions) them gives a better result ('conversion') than the alternative. This led

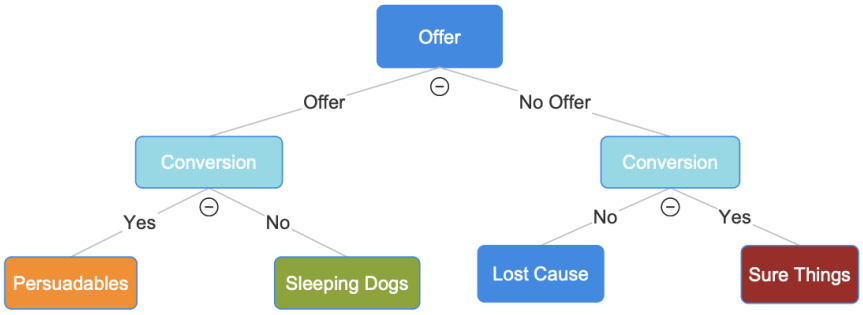


Figure 1: Decision tree for determining customer type

us to create a binary variable called '**treat**', where 0 indicates not sending a promotion if the customer type is not 'Persuadables' and 1 if it is, that would serve as our target variable. After introducing this new variable, we saw that the distribution of the target variable was unbalanced as we had: 1: 42694, 0: 21306. To address this issue, we implemented SMOTE algorithm to balance the distribution by oversampling. Finally, we had an equal distribution of the 'treat' variable of 1: 42614, 0: 42614.

On the dataset, a mapping was applied to the categorical features which had a string type. As for the "offer" column, the 'condition' function assigns a value of 1 for "Buy One Get One," -1 for "Discount," and 0 for any other offer types. In the "channel" column, the mapping is such that "Web" is encoded as 1, "Phone" as 2, and any other channels as 3. Similarly, the "zipcode" column is mapped with "Suburban" given a value of 1, "Urban" a value of 2, and other types of regions a value of 3.

3 Black Box Model Application

We aimed to predict the feature 'treat' from the other features of our dataset, excluding 'offer', 'conversion', 'treat', and 'customer_type' using three different black box models: Random Forest, AdaBoost, and Gradient Boosting.

By using three different black-box models we wanted to investigate how each model reacts to the data, as well as their robustness. For each model, we applied a 7-fold cross-validation to find the best hyperparameters, and we obtained the following results and hyperparameters:

1. Gradient Boosting: Achieved the highest accuracy (79.3%) with parameters learning_rate:0.5, max_depth:10, subsample:1, n_estimators: 75 and the others set to default.
2. Random Forest: Had a lower accuracy (72.5%) with parameters max_depth:10, n_estimators:50, max_features:5 and the others set to default.
3. AdaBoost: Showed the lowest accuracy (66.4%) with parameters learning_rate:0.9, n_estimators:15 and the others set to default.

Based on these results, we decided to continue with Gradient Boosting classifier, whose detailed performance can be seen in Figure 2, as our black box model for which we find and analyze explanations in the following section.

```
Best parameters: {'learning_rate': 0.5, 'max_depth': 10, 'n_estimators': 75, 'subsample': 1}
Accuracy 0.7933073637771625
F1-score [0.78833029 0.79805576]
              precision    recall  f1-score   support

           0       0.81      0.77      0.79     10653
           1       0.78      0.82      0.80     10654

    accuracy                           0.79     21307
   macro avg       0.79      0.79      0.79     21307
weighted avg       0.79      0.79      0.79     21307
```

Figure 2: Gradient Boosting Classifier Parameters

4 Explainability Methods

In this report, we explore the use of three prominent XAI techniques: Local Interpretable Model-Agnostic Explanations (LIME), Layer-wise Relevance Propagation (LORE), and SHapley Additive exPlanations (SHAP). By employing this diverse set of methods, we aim to achieve a richer understanding of the factors influencing the model's predictions. This in order to fulfill the needs of various stakeholders, including clients, users, colleagues, and developers, by providing explanations tailored to their technical expertise.

For our implementation, we utilized the **XAI-LIB** library for LORE explanations and the dedicated **lime** and **shap** libraries for LIME and SHAP, respectively. These choices ensured optimal support for generating explanations specific to individual data instances. The explanations primarily focus on observation 147 (unless otherwise specified) obtained from the test set.

4.1 Evaluation of Explainability Methods

4.1.1 Content

Correctness/comprehensiveness refers to whether the explanation captures all relevant factors influencing the model's prediction of whether a customer should be treated or not. In this context, a correct and comprehensive explanation might highlight: **Purchase history**: High historical spending could indicate a valuable customer who might benefit from treatment. Conversely, low spending might suggest treatment wouldn't be effective. **Acquisition channel**: Customers acquired through referrals might be more receptive to treatment, while those acquired through generic channels might require stronger justification. **Recency**: Recent purchases could indicate higher customer engagement, making them good candidates for treatment. **Past promotions**: Customers who used buy-one-get-one offers or discounts might be more price-sensitive and less likely to respond to additional treatment. **Zip code**: The customer's zip code classification (Suburban/Urban/Rural) could influence treatment decisions. For example, the model might prioritize treatment for customers in areas with lower overall spending or less access to similar products.

Completeness/sufficiency focuses on whether the highlighted features in the explanation are enough to explain the model's decision fully. Here, we want to ensure the explanation doesn't omit crucial factors: **Is referral**: While the explanation might focus on the acquisition channel (referral vs. non-referral), a complete explanation should clarify if the specific referral status (yes/no) played a role in the treatment decision. By analyzing both **correctness** and **completeness**, we can gain valuable insights into the transparency and reliability of the XAI method and the explanations it provides for customer treatment decisions."

Continuity essentially implies similar Input with have similar explanations. This principle emphasizes that explanations for similar customer profiles should be comparable. In our context, customers with analogous characteristics (purchase history, acquisition channel, recent purchases, past promotion usage, and zip code classification) should receive explanations that reflect these similarities. For example, if two customers have a high purchase history and haven't used discounts recently, the XAI should explain why treatment might be beneficial for both, potentially highlighting their loyalty and potential value.

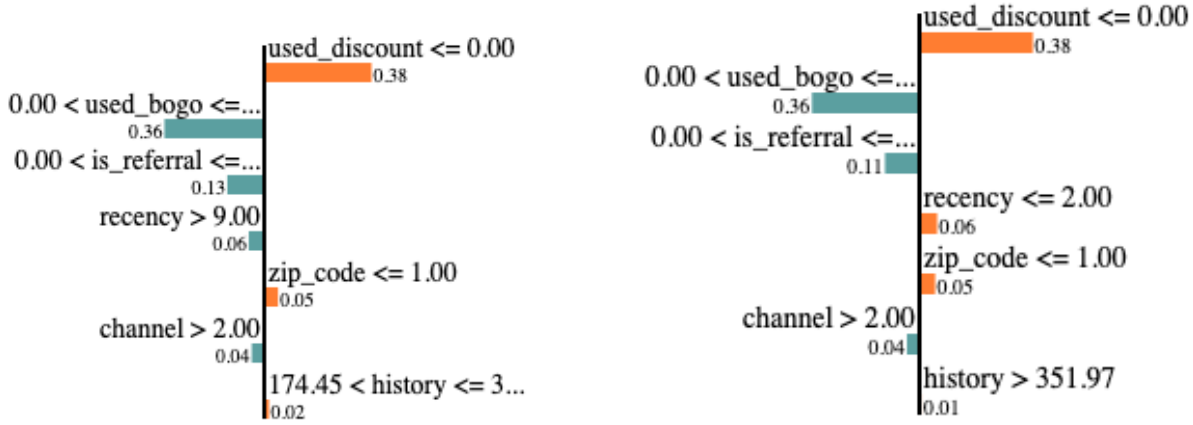
To highlight this we used instance 792 and instance 10301, which have these values in their attributes: **Instance 792** -> channel:3,history:299.45,is_referral:1,recency:11,used_bogo:1,used_discount:0,zip_code:1

For **Instance 10301** -> channel:3,history:404.69,is_referral:1,recency:1,used_bogo:1,used_discount:0,zip_code:1

In the **LIME evaluation** is possible to see how the importance to classify both observations reflect a similar composition this shows continuity on the explainable method, it's also worth noting how for **SHAP** is possible to see **use_discount** as the only relevant one for contributing to **no treatment** in both, in **instance 792 recency** is connected with **treatment** but the level of importance is one of the lower, and for **instance 10301 recency** also is shown as less important for classification as **no treatment**.

(a) Instance 792 (b) Instance 10301

Figure 3: SHAP evaluation of continuity



(a) Instance 792

(b) Instance 10301

Figure 4: LIME evaluation of continuity

4.1.2 Presentation

Compactness refers to the quality of the explanations being concise, brief, and clear to understand without losing important information. To determine the compactness of our explainers, we calculated the following information: For SHAP, we evaluated the number of non-zero SHAP values we had for our specific instance. A non-zero value indicates that a feature contributes to the prediction, and in our case, we only had 1 non-zero value. This means that the explanation is very compact, highlighting that a single feature is predominantly driving our model's prediction for this specific instance. For LIME, we evaluated the number of features in the explanation, to see which attributes influenced the decision making. The value we obtained was 7 features, which shows relative compactness, though fewer features would show a higher level of model conciseness and compactness. For LORE, we evaluated the length of the rules we got from the explanation. Our longest rule set contained 11 rules, whereas our smallest rule contained 2 rules. This again shows relative compactness. We also found the mean, median and mode value for the length of the rules and we found a consistent result of 6 in all those metrics. It is important to highlight that the metrics were calculated on a total of 10,000 explanations.

Composition refers to defining the presentation format, organization, and structure of the explanation. To determine the composition of our explainers, we take a look at the following information: For SHAP, we see in 5a the use of a beeswarm plot to determine the contribution of specific instances. For LIME, the plots represent the explanations in the form of a set of features and their corresponding contributions. As we can see in 5b, similarly the beeswarm plot shows the positive and negative influence of attributes in our explanations. For LORE, the explanation is given in a different way, in the form of a decision tree rules where we explain the range of every attribute that affected the result to be as such. An example can be seen in 5c.

4.1.3 User

To evaluate coherence, We have chosen to analyze a randomly selected instance characterized by the following attributes:

- **Channel**: 1.0 (Web)
- **History**: 193.9584149005614 (historical purchase value)
- **Is Referral**: 0.0 (not a referral)
- **Recency**: 1.0 (1 month since the last purchase)
- **Used BOGO**: 1.0 (utilized buy one get one offer)
- **Used Discount**: 0.0 (did not use a discount)
- **ZIP Code**: 2.0 (Urban)

True Class: 1 (target for ads)
Predicted Class: 1 (target for ads)

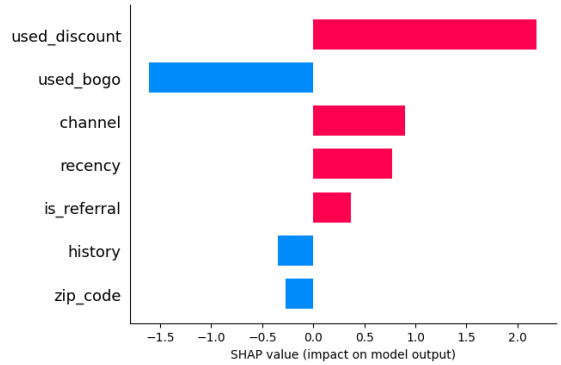
Given that this customer made a purchase online, their historical purchase value is close to the median (174), they did not use the referral program, they made a recent purchase, used a Buy One Get One offer, did not use a discount, and reside in an urban area, we can proceed with analyzing the results from various explainers.

From a qualitative perspective, we consider the following factors favorable:

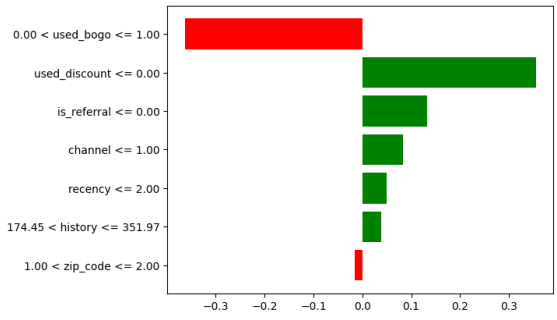
- Channel: Most modern advertising is conducted online.
- History: The historical purchase value aligns with the median value of other customers in the dataset.
- Used Discount: The lack of discount usage might indicate a higher spending capacity.

Conversely, an unfavorable factor is the recent purchase (one month ago), which might suggest less immediate need for additional purchases.

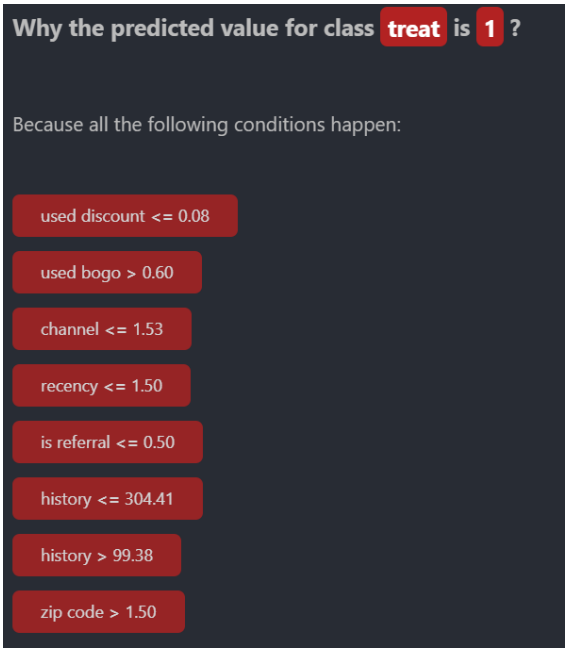
Using SHAP, as illustrated in Figure 5a, we observe that the explanation aligns with our qualitative assessment. Interestingly, the ZIP code, which was, in fact, not initially considered, appears as the least significant factor in the explanation.



(a) SHAP Instance Explanation



(b) LIME Instance Explanation



(c) LORE Instance Explanation

Similarly, LIME results, shown in Figure 5b, indicate that the BOGO offer is the most significant decremental factor. The rest of the explanation is consistent with our qualitative assessment.

Finally, LORE, depicted in Figure 5c, provides a direct explanation for the predicted class of 1, confirming the reasons identified by LIME and aligning with our qualitative assessment.

To evaluate Context, the explanation provided is highly relevant to the user and their needs for several reasons:

- **Contextual Relevance**: The explanation focuses into specific attributes of a randomly selected customer instance, such as purchase history, recency, usage of offers, and demographic details. This context is crucial for users looking to understand the decision-making process behind targeting specific customers for advertisements.
- **Qualitative Assessment**: The analysis offers a qualitative assessment, aligning customer attributes with favorable and unfavorable factors for targeting ads. This helps users grasp the practical implications of these attributes in real-world scenarios.
- **Alignment with Predictive Models**: The explanation uses SHAP, LIME, and LORE to cross-validate the findings. By showing almost **complete** consistency across multiple models, it enhances the credibility of the results, which is vital for users needing robust and reliable explanations.
- **Clear Visual Aids**: Figures illustrating SHAP, LIME, and LORE instance explanations provide visual clarity, helping users to easily interpret the significance of different factors. This visual support is beneficial for users who may prefer or require graphical representations to understand complex data.
- **Insight into Predictive Factors**: The explanation highlights which factors are most and least significant in predicting the target class. This insight is critical for users aiming to refine their strategies or understand which customer attributes influence the targeting decisions the most.

Overall, the detailed, multi-faceted analysis caters well to users' needs for comprehending how and why specific customers are targeted, providing both qualitative insights and quantitative validations through multiple explanatory models.

5 Research Questions

5.1 How can explanations help in finding the cases when black-box is "right for the wrong reasons"?

To investigate instances where a black-box model makes correct predictions based on potentially incorrect reasoning, we implemented three explainable AI (XAI) methods: LIME, SHAP, and LORE. These were applied to a gradient boosting classifier.

Our approach consisted of the following steps:

1. Loaded a pre-trained gradient boosting model and associated train/test datasets.
2. Selected 10,000 correctly classified instances for analysis.
3. For each XAI method, create a dictionary of the most relevant features and intervals, along with their frequency across the sample set.
4. Identified the top features and analyzed correctly classified instances that did not rely on these primary features.

Initially, we considered the top 7 features for each method. However, we found that the most relevant features were often within the top 3 of each explanation. Therefore, we focused on the top 3 features for a more precise analysis.

LIME and SHAP yielded identical top features:

- is_referral <= 0.00 (frequency: 5750)
- used_discount <= 0.00 (frequency: 5180)
- channel <= 1.00 (frequency: 5086)
- 0.00 < used_bogo <= 1.00 (frequency: 5071)
- zip_code <= 1.00 (frequency: 4962)
- used_bogo <= 0.00 (frequency: 4929)
- 0.00 < used_discount <= 1.00 (frequency: 4820)

LORE analysis was conducted on a smaller sample (250 instances) due to computational constraints. The top features identified were:

- history (frequency: 395)

- recency (frequency: 290)
- used_discount (frequency: 207)
- channel (frequency: 167)
- used_bogo (frequency: 166)
- zip_code (frequency: 144)
- is_referral (frequency: 138)

We iteratively reduced the number of top features considered to identify how many instances were classified correctly but potentially for the wrong reasons. The results are presented in Table 1.

Table 1: Number of Potentially Misclassified Instances by Feature Set Size

Method	Top 7	Top 6	Top 5	Top 4	Top 3	Top 2
LIME	0	0	3982	3982	4820	4820
SHAP	0	0	3982	3982	4820	4820
LORE	0	0	0	0	0	0

When all top 7 features are considered, no instances are classified as potentially incorrect. However, as we reduce the feature set, we observe an increase in potentially misclassified instances. For example, when considering only the top 3 features for LIME and SHAP, 4,820 instances are correctly classified but potentially for the wrong reasons, as their explanations do not include the features "zip_code <= 1.00", "used_bogo <= 0.00", or "0.00 < used_discount <= 1.00".

Regarding LORE, we were unable to extract this data since the rules generated are not always more than one. This limitation prevented us from conducting the same type of analysis performed with the other two methods. However, we can still infer information from the most common rules previously described.

This analysis demonstrates the importance of comprehensive feature consideration in explaining model decisions and highlights potential areas where the model may be relying on less relevant features for classification.

5.2 How does presenting different explanations support usability?

In itself, usability refers to the degree of ease that users can effectively interact with and understand a model. As mentioned before, we have already applied three different types of explanations: LIME, LORE, and SHAP. Our reasoning to do so aligns with the question, as we believe that presenting a diverse range of explainers can help users feel at ease and trust the decisions of a model. More specificilly, they offer:

1. Broad range of accessibility
2. Diverse Visualization Properties
3. Different Aspects of Error Detection

Firstly, for accessibility, a broader ranger of explainers allows us to increase accessibility for various stakeholders, technical and non-technical alike. As such, SHAP provides a two dimensional information: simple summary plots and feature importance for more non-technical users, and SHAP values rooted in game theory for more technical ones. Moreover, LIME is more suited for non-technical users who through the plots can be able to see the probabilities for an instance being in a particular class, and what features influence that decision. LORE provides rules which identify with both non and technical users who like to have a logic walk-through into the decision making.

In regards to visualization, LORE provides rule-based visualizations can be very interpretable (unless the rules become too complex), especially for users familiar with decision rules like data engineers or technology expertes. SHAP, on the other hand, provides summary plots, dependence plots, and force plots which show the global/local importance of features and interactions in a more simplified and aesthetic way. LIME's visualizations include bar charts that show feature contributions.

As for error detection, the different visualization properties can help us identify diverse aspects of errors that our model might be making. As such, in SHAP we are able to identify biases on our model when seeing which features contribute more to the prediction and evaluating if they're coherent with human thinking. LIME that usually handles local explanations can help identify anomalies in our data, whereas LORE which gives us decision rules that can help identify any logical incosistency in the model.

Ultimately, having more than one explainers supports usability by offering more choice in visualization, a more in-depth understanding of errors, as well as tailoring to different stakeholders.

5.3 How do different explanations support users with different levels of expertise?

In the case of having a model predicting uplift modeling for marketing campaigns, there exist various stakeholders might interact with the dataset and the model. As an example, we will see the following three stakeholders and their respective levels of expertise:

1. Executives -> Beginner
2. Marketing Agents -> Intermediate
3. Data Engineers -> Advanced

Starting with executives, their role is to oversee the overall marketing strategy and making very high-level business decisions. As such, an explainer like SHAP that offers clear and intuitive visual representations of feature importance through its beeswarm plot, would be an ideal choice for executives. In the beeswarm plot we can tell the key driving features that determine whether or not a person would buy a product through its simple presentation and bold colors, without delving into too much technical information at first glance. This simplicity in visual presentation allows executives to quickly grasp the information without first requiring to have a deep technical knowledge.

For marketing agents, who want to have a bit more specific insight, we think an explainer like LIME would be most suitable. The reason being that LIME helps understand how specific customers audience and their characteristics can influence the outcomes of their campaigns. For example, our LIME results show us that if a user used a discount, they are more prone to buying a product. This can effectively influence marketing strategies, which might focus more now on devising and distributing discount codes/emails to their customers.

Finally, our most advanced users would be the data engineers whose job is to ensure and evaluate the accuracy and fairness of the model. To make sure that the model is giving correct results for the "right reasons" or understanding the behind logic when it gives incorrect results, a data engineer would most benefit from an explainer like LORE. Since LORE provides rule-based explanations, which give information about how specific attribute values and their ranges influence model predictions, it helps data engineers in two dimensions: debugging, which would help verify if the model's logic is performing as expected, as well as error detection: insights for instances can help to identify any anomalies or biases in the model.

These different users determinately show that having diverse explainers allows us to cater to a broader audience of users, and effectively explain to them how a certain model "thinks", but always giving the amount of information that is necessary to each user. As we noted, an executive wants a more high-level view of information and benefits from simplicity, a marketing agent wants a bit more detailed view into specific customers and benefits from a local explainer, while a data engineer wants a more in-depth understanding of the inner workings of the model.

6 Conclusion

In conclusion, this report demonstrates the critical role of Explainable AI (XAI) in making black-box models comprehensible and trustworthy. By implementing LIME, SHAP, and LORE on a gradient boosting model for uplift modeling, we provided clear, accessible explanations tailored to various kinds of users. The study highlighted how each explainer aids different users: SHAP offers intuitive visual summaries ideal for executives, LIME provides actionable insights for marketing agents, and LORE delivers detailed rule-based explanations suitable for data engineers.

The analysis revealed that a comprehensive approach to explainability enhances model transparency, supports error detection, and accommodates diverse expertise levels. SHAP and LIME identified key predictive features, while LORE's rule-based explanations ensured logical consistency. This multi-faceted approach ensures that explanations are both contextually relevant and technically robust, promoting confidence in AI-driven decisions across different user groups.

Overall, the findings underscore the importance of using diverse XAI methods to meet the varied needs of users, enhancing both the usability and reliability of AI models. By doing so, organizations can better understand, trust, and effectively leverage AI in decision-making processes, ultimately leading to more informed and strategic outcomes.