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 '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 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 gives a better result ('conversion') than the alternative. This lead 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

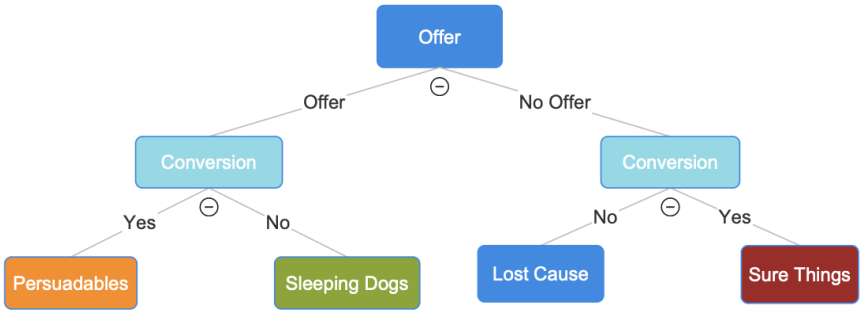

Figure 1: Decision tree for determining customer type

implemented SMOTE algorithm to balance the distribution by oversampling. Finally, we had an equal distribution of the 'treat' variable of 1: 42614, 0: 42614.

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

Based on these results, we decided to continue with Gradient Boosting classifier as our black box model for which we find and analyze explanations in the following section.

4 Explainability Methods

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 (Get advertised) 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 fully explain the model's decision. Here, we want to ensure the explanation doesn't omit crucial factors:

Is referral: While the explanation might focus on 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** Similar Inputs, 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.

PUT THE RESULTS XAI

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.

5.1.1 Methodology

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, created 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.

5.1.2 Results

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)

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

5.1.3 Analysis of Misclassifications

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.

5.1.4 Interpretation

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 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 level 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 LORE would be most situatable. The reason being that LIME helps understand how specific customer attributes, for example a specifc influence the outcomes of their campaigns.

6 Conclusion