DOI: 10.21917/ijsc.2022.0396

COMPARATIVE STUDY OF XAI USING FORMAL CONCEPT LATTICE AND LIME

Bhaskaran Venkatsubramaniam and Pallav Kumar Baruah

Master of Math and Computer Science, Sri Sathya Sai Institute of Higher Learning, India

Abstract

Local Interpretable Model Agnostic Explanation (LIME) is a technique to explain a black box machine learning model using a surrogate model approach. While this technique is very popular, inherent to its approach, explanations are generated from the surrogate model and not directly from the black box model. In sensitive domains like healthcare, this need not be acceptable as trustworthy. These techniques also assume that features are independent and provide feature weights of the surrogate linear model as feature importance. In real life datasets, features may be dependent and a combination of a set of features with their specific values can be the deciding factor rather than individual feature importance. They also generate random instances around the point of interest to fit the surrogate model. These random instances need not be part of the original source or may even turn out to be meaningless. In this work, we compare LIME to explanations from the formal concept lattice. This does not use a surrogate model but a deterministic approach by generating synthetic data that respects implications in the original dataset and not randomly generating it. It obtains crucial feature combinations with their values as decision factors without presuming dependence or independence of features. Its explanations not only cover the point of interest but also global explanation of the model, similar and contrastive examples around the point of interest. The explanations are textual and hence easier to comprehend than comprehending weights of a surrogate linear model to understand the black box model.

Keywords:

Explainable AI, XAI, Formal Concept Analysis, Lattice for XAI, Deterministic methods for XAI

1. INTRODUCTION

It is very common to use a Machine Learning or a Deep Learning model to be trained on a dataset and later in production as the model peaks in evaluation parameters, say accuracy. While these models excel at fitting the curve to data, barring a few, most of them cannot explain why a specific decision was made. The model learns patterns from training data and predicts an outcome for an instance or maps an instance to a class. In a black box model, the learnt patterns of data are not evident and the reasons why a model decided an outcome is not clear. In order to make these models trustworthy and therefore acceptable, it is necessary to augment the model with explanations of its decisions. One approach is to drop these black box models and adopt only white box models [1]. But that may be an extreme approach as many Deep Learning models are here to stay. Yet it may be foolish to adopt these in production without knowing the reason behind their decision and hence explainable AI becomes a necessity [2].

Local Interpretable Model Agnostic Explanation (LIME) is a very popular technique to extract explanations from a black box model [3]. It generates local explanation at a point of interest by generating random data around it, finding the model outcome of these data instances and by fitting a linear model to this randomly generated data. It uses the weights of the surrogate linear model to be presented as the local explanation at the point of interest. While this is a good technique to understand the model behavior locally at a point of interest, it need not be a robust method of explainability [4]. Similar data instances may not have similar explanations challenging the very purpose of explainability. LIME also faces challenges of stability in producing the same explanations employing repeated use of the method under the same conditions [5]. Lack of robustness or instability in these methods challenge their applicability, specifically in sensitive domains like healthcare. While linear models consider each feature independent from the other, it may not be inherently true in the dataset. Approximating the black box model with a linear model can be locally faithful but may not be so in a larger perspective. Using an aggregate of the weights of linear models around representative data instances cannot substitute for global explanation [6]. Some of these issues are prevalent in other methods of explainability too.

Other approaches in explainable AI use an unified approach [7], or class activation maps for image data to produce saliency visualizations [8]-[10]. Visualizations are good at generating intuition about an instance but there are no rigorous validations to extrapolate these intuitions to the entire model [11]. While heat map visuals depict where the model is looking, most often human intuition fills up what the model is looking at. Many techniques indicate the reasons why an instance belongs to a specific class, while not presenting reasons for the instance not being in another class. Presenting a heat map or weights as explanations needs user expertise to understand the explanations themselves [1]. Some of these techniques fail invariance to another equivalent implementation or non-intrusive input transformations [12] and some are not sensitive to model parameter or model-outcome relationship randomization [13].

In this work, we compare our novel technique to extract explanation from the lattice [14] to that of LIME. Lattice based explanation extracts several combinations of features and their values responsible for an outcome retention or change. It not only extracts a set of feature implications to different classes acting as global explanation, rightfully representing the black box model, but also produces a hierarchy of minimal feature combinations of an instance that lead to the model's decision of a specific class and not other classes, acting as local explanation. Apart from global and local explanations, this technique also provides similar and contrastive explanations around an instance. Feature combinations that distinguish an instance away from other classes are considered salient while feature combinations that do not contribute to this distinction do not appear in the explanation. Implications gathered from the dataset and the model outcome act as rigorous validations enabling extrapolation of the global explanation to the entire model. Meaningful relationships in the dataset are utilized to build the lattice on realistic synthetically generated data instead of unrealistic random data.

In similar works that use a lattice for explanation, [15] use the lattice to guide samples chosen by LIME, while [16] build a lattice-based model independent of the black box model and compare it with the classification model to produce only individual features responsible for the decision. In this work, we compare the explanation from the lattice [14] to explanations from LIME.

Section 2 introduces the formal concept lattice and the overview of our novel technique to extract explanations from the lattice [14]. While the assumption of a linear model by LIME is good for interpreting instances locally, it does not provide a global intuition. We use a popular UCI dataset [17] to prove this limitation of LIME and compare it to the global explanation from the lattice in Section 3. LIME assumes each feature to be independent and presents the weight of the feature as its importance. But it is possible that a combination of a set of features with their specific values were responsible for determining the outcome. We use the UCI dataset to prove this limitation of LIME by comparing its local explanations to that from the lattice in Section 4. LIME generates many random data instances around the point of interest to fit a linear model at the point of interest. These random data instances do not follow any implications in the given dataset and may turn out to be a meaningless combination in the real dataset. Moreover, these data instances are generated for each point of interest. In lattice-based explanations, we generate synthetic data globally based on implications from the dataset and user provided implication cutoff, as stated in [14]. Section 6 contains conclusions and future work.

2. FORMAL CONCEPT LATTICE

A context is a triple (G,M,I), where G is a set of objects, M is a set of attributes and I the relation between them. The notation *gIm* means that the object g has the attribute m.

For a set $A \subseteq G$, define $A' = \{m \epsilon M \mid gIm \ \forall \ g \epsilon A\}$ [A' is the set of attributes common to all the objects in A]

For a set $B \subseteq M$, define $B' = \{g \epsilon G \mid gIm \ \forall \ m \epsilon B\}$ [B' is the set of objects which have all attributes in B]

A concept of the context (G,M.I) is a pair (A,B) such that, $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. A is called the extent and B the intent of the concept (A,B).

If (A_1,B_1) and (A_2,B_2) are concepts of a context (G,M,I), then (A_1,B_1) is a subconcept of (A_2,B_2) (or (A_1,B_1) is a superconcept of (A_2,B_2)), denoted by $(A_1,B_1) \leq (A_2,B_2)$ (or $(A_2,B_2) \leq (A_1,B_1)$) if $A_1 \subseteq A_2$, equivalently $B_2 \subseteq B_1$ (or $A_2 \subseteq A_1$, equivalently $B_1 \subseteq B_2$). The relation $\leq$ is called the hierarchical order of the concepts. The ordered set of concepts is called the concept lattice of the context (G,M,I). Concept lattices are represented using a hasse line diagram [18]. The Table.1 contains a simple formal context and Fig.1, its concept lattice.

Table.1. Formal context of a few species with their attributes (columns split into two parts)

	Breathes in water (a)	**Can fly (b)**	**Has beak (c)**	**Has hands (d)**	**Has skeleton (e)**
Bat		X			X
Eagle		X	X		X
Monkey				X	X
Parrot Fish	X		X		X
Penguin			X		X
Shark	X				X
Lantern Fish	X				X

	Has wings (f)	**Lives in water (g)**	**Is viviparous (h)**	**Produces light (i)**
Bat	X		X	
Eagle	X			
Monkey			X	
Parrot Fish		X		
Penguin	X	X		
Shark		X		
Lantern Fish		X		X

In [14], we use the formal concept lattice to extract global, local, similar and contrastive explanations of a black box model around an instance of interest. From the given dataset, a formal concept lattice is temporarily constructed to derive a set of implications. Based on the user's implication cutoff, a synthetic dataset is created respecting the implications whose support is greater than or equal to the cutoff. Using this synthetic dataset and the model outcome, a formal concept lattice is constructed and implications from this are presented as the global explanation. For local, similar and contrastive explanation, this lattice is traversed to find minimum feature combinations that lead to a specific outcome. Multiple sanity tests and comparison with a white box model prove its credibility.

3. GLOBAL EXPLANATION COMPARISON BETWEEN LIME AND LATTICE ON A UCI DATASET

LIME provides a global understanding of the model using the submodular pick module [3]. Based on the budget factor (time/patience), a certain number of instances are picked and explanations of these instances are combined to form a global explanation. Intuitively, global importance of features are higher for those that explain multiple instances. The instances are picked in a non-redundant manner ensuring that each instance adds value to the explanation.

While it is true that a set of picked instances can provide a general trend, it need not be the true perspective at all the instances and such lack of accuracy in explanations may prove costly. At the same time, budget constraints may not allow for all the instances to be covered. Even if all instances are covered, an aggregate of all the weights of features may not indicate the global picture due to the fundamental assumption of linearity.

Our approach using the lattice to generate explanations differs from LIME. We generate synthetic data based on the implications derived from the dataset and user provided implication cutoff (similar to budget). This type of synthetic data is not random (as generated by LIME) and is more meaningful to the real world data source as it respects the implications in the original dataset. With this synthetic dataset and the model outcome, a lattice is constructed from which different types of explanations are extracted [14]. This is advantageous over LIME as the lattice is built on the entire dataset following the specified implications thereby providing the true global perspective instead of relying on an aggregate of specific individual instances.

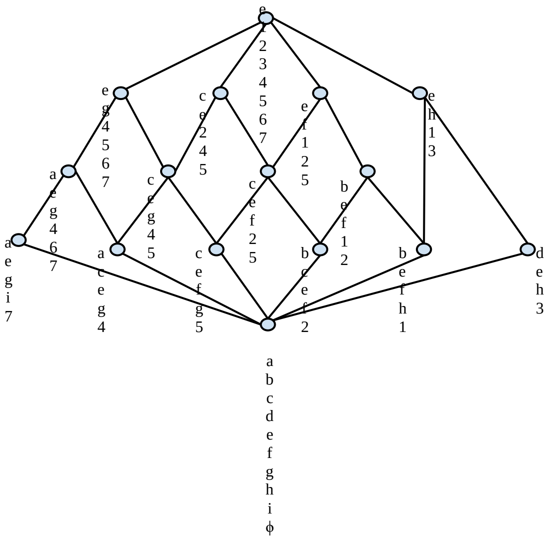


Fig.1. Concept Lattice of the formal context in Table.1

In this section, we compare the global explanations of a black box model for the UCI Car Evaluation dataset [19] between lattice and LIME. The Car dataset has the following features:

buying: vhigh, high, med, low [Indicates purchase price of the car]

maint: vhigh, high, med, low [Indicates maintenance cost of the car]

doors: 2, 3, 4, 5-more [Indicates number of doors in the car]

persons: 2, 4, more [Indicates number of persons that the car can carry]

lug_boot: small, med, big [Indicates the luggage and boot space]

safety: low, med, high [Indicates the level of safety built in the car]

class: acc, good, unacc, vgood [Indicates if such a car is acceptable, good, unacceptable or very good]

Feature string values were converted to numeric type and a Random Forest classifier was trained with 80% of this data that delivered 96.5% accuracy.

3.1 GLOBAL EXPLANATION FROM LIME

The Global explanation by submodular pick LIME was generated with the entire dataset and 5 data instances for representing the global explanations (SP-LIME method='full', num_exps_desired=5). We present the explanation of these five instances below.

a) Instance features: *buying=high, maint=low, doors=2, persons=2, lug_boot=big, safety=low*

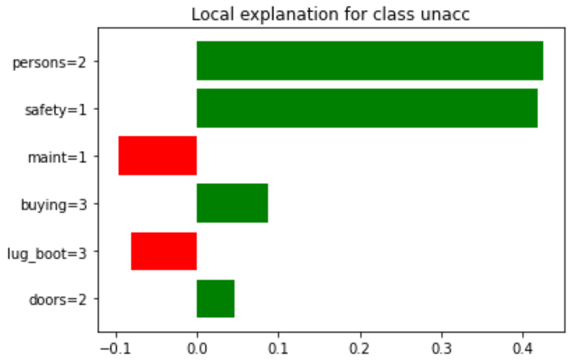


Fig.2. LIME explanation of the first instance

The explanation in Fig.2 states that since the car can carry only two persons and has low safety, they were primarily responsible to classify this instance unacceptable Even with low maintenance cost and a large luggage and boot space, this instance could not be acceptable as the weights of the two primary features were comparatively large.

b) Instance features: *buying=med, maint=high, doors=5more, persons=4, lug_boot=med, safety=high*

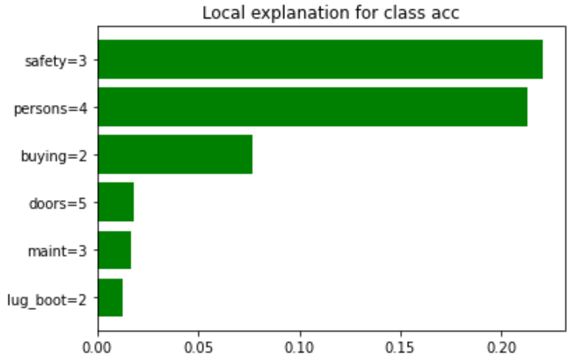


Fig.3. LIME explanation of the second instance

The explanation in Fig.3 states that since the car has high safety and can carry 4 persons, they are primarily responsible for it being acceptable, while other features also add value towards the decision.

c) Instance features: *buying=low, maint=med, doors=4, persons=more, lug_boot=small, safety=med*

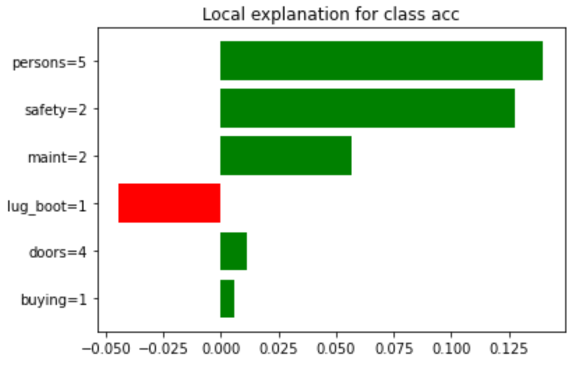


Fig.4. LIME explanation of the third instance

The explanation in Fig.4 states that since the car can carry more than 4 persons, has medium safety and a low maintenance cost, despite having a small luggage and boot space, it is acceptable

d) Instance Features: *buying=vhigh, maint=vhigh, doors=3, persons=4, lug_boot=big, safety=high*

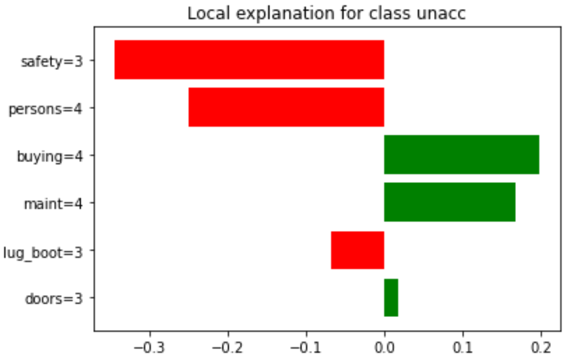


Fig.5. LIME explanation of the fourth instance

The explanation in Fig.5 states that despite the car having a high safety and carrying four persons, since the buying and maintenance costs are very high, it is unacceptable

e) Instance Features: *buying=med, maint=high, doors=5more, persons=2, lug_boot=big, safety=high*

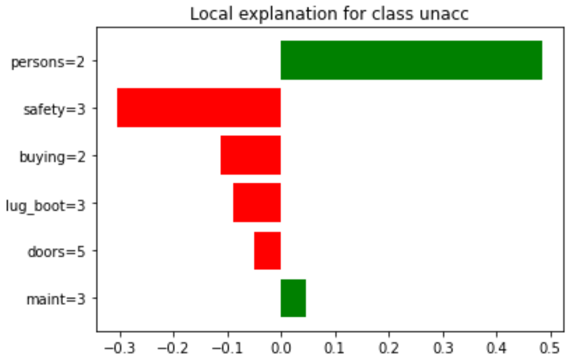


Fig.6. LIME explanation of the fifth instance

The explanation in Fig.6 states that despite the car having high safety, medium buying cost and a big luggage and boot space, since it can carry only two persons, it is unacceptable

The Global explanation from LIME is only indicative using representative data instances and does not provide a larger picture of the model. Considering the five instances above, there is no clarity of what the model could do for another data instance.

3.2 GLOBAL EXPLANATION FROM LATTICE

Global explanation was generated from the Lattice respecting all implications in the entire dataset (cutoff=0). The explanation from the lattice is in the form of feature value combinations leading to a specific class or a set of classes. This form is much more intuitive to understand compared to weights of features as presented by LIME. More than that, these implications provide a true global understanding of the model. Few of these are as below:

1. (buying,high) ⇒ acc or unacc

2. (buying,vhigh) ⇒ acc or unacc

3. (maint,high) ⇒ acc or unacc or vgood

4. (maint,vhigh) ⇒ acc or unacc

5. (persons,2) ⇒ unacc

6. (lug_boot,small) ⇒ acc or good or unacc

7. (safety,low) ⇒ unacc

8. (safety,med) ⇒ acc or good or unacc

9. (buying,high)(maint,vhigh) ⇒ unacc

10. (buying,med)(maint,high) ⇒ acc or unacc

11. (buying,med)(maint,med) ⇒ acc or unacc or vgood

12. (buying,vhigh)(maint,high) ⇒ unacc

13. (buying,vhigh)(maint,vhigh) ⇒ unacc

14. (maint,high)(lug_boot,small) ⇒ acc or unacc

15. (maint,high)(safety,med) ⇒ acc or unacc

16. (doors,2)(lug_boot,med) ⇒ acc or good or unacc

17. (lug_boot,big)(safety,high) ⇒ acc or unacc or vgood

18. (lug_boot,small)(safety,med) ⇒ acc or unacc

These are very simple and easy to understand. Implication no. 5 states that if a car is capable of carrying only two persons it is declared to be unacceptable by the model despite having the best values for other features. Similarly implication no. 7 states that a car is unacceptable if its safety is low and implication no. 8 states that a car can never be classified as very good with medium safety. We verify these implications by setting the best feature values except for the ones stated in the implication and pass it to the model.

Instance Features for implication no. 5: *buying=low, maint=low, doors=5more, persons=2, lug_boot=big, safety=high*

Model outcome: [0.0, 0.0, 0.988, 0.012] classifying it as unacceptable as stated in the implication.

Instance Features for implication no. 7: *buying=low, maint=low, doors=5more, persons=more, lug_boot=big, safety=low*

Model outcome: [0.0, 0.028, 0.966, 0.006] classifying it as unacceptable as stated in the implication.

Instance Features for implication no. 8: *buying=low, maint=low, doors=5more, persons=more, lug_boot=big, safety=med*

Model outcome: [0.026, 0.938, 0.012, 0.024] classifying it as good, but not very good, as stated in this implication.

This clearly proves that explanation from the lattice is much more acceptable as a global understanding of the model than an aggregate of multiple local explanations at different data instances.

4. LOCAL EXPLANATION COMPARISON BETWEEN LIME AND LATTICE ON A UCI DATASET

In this section, we continue to use the Car Evaluation dataset to compare the explanations of a black box model from the lattice and LIME for data instances belonging to different classes.

4.1 INSTANCE FROM UNACCEPTABLE CLASS

Instance (a): *buying=vhigh, maint=med, doors=3, persons=4, lug_boot=small, safety=low*

The Random Forest model prediction probabilities for this data instance are [0.0, 0.0, 1.0, 0.0] clearly classifying it as unacceptable.

4.1.1 Explanation from LIME:

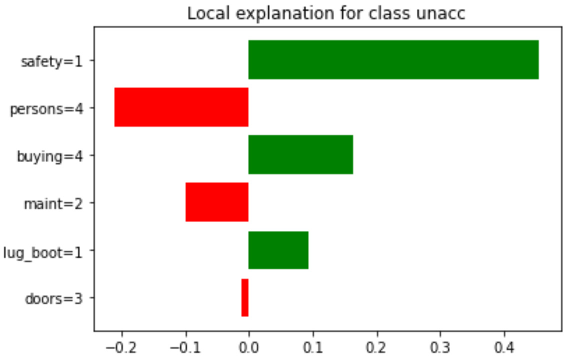


Fig.7. LIME explanation of instance (a)

In Fig.7, LIME identifies that the feature safety being low, pulls the car to the unacceptable class with a large weight, despite the car being capable of carrying four persons and running at a low maintenance cost, that tried to take it away from the unacceptable class. The next factors that caused the car to be unacceptable are a very high buying cost and a small luggage and boot space.

4.1.2 Local Explanation from the Lattice:

The following local explanation is obtained for instance (a) from the lattice:

Features: (buying,vhigh) deny class(es): good vgood

Remaining class(es) are: acc unacc

Features: (safety,low) deny class(es): acc

Remaining class(es) are: unacc

Lattice traversal has denied all class(es) except unacc

Features: (safety,low) lead to class(es) unacc with a confidence of 1.000000

The lattice explanation not only provides those features that lead a car to a specific class, but also the features that denied other classes. Since the buying cost is very high it cannot belong to the good or very good category. Added to it since its safety is low it cannot be acceptable either. It uses the same feature of safety being low to make it unacceptable with a confidence of 1.0. This matches with the highest weight feature as identified by LIME.

But there are also mismatches between the two explanations. While LIME indicates that buying cost, luggage and boot space positively influence the decision of the model, explanation from the lattice does not include these features at all. In order to check if it is an anomaly from the lattice and verify the truth, we modify instance (a) with the best values for buying cost, luggage and boot space and provide it as input to the model.

Instance (a'): ***buying=low***, *maint=med, doors=3, persons=4,* ***lug_boot=big***, *safety=low*, where, we have set the buying cost to be low (the least of all values) and luggage and boot space to be big (largest of all values) keeping the rest of the features the same. If the influence of buying cost, luggage and boot space is large enough, then the class should change.

But the Random Forest model prediction probabilities for this data instance are [0.0, 0.012, 0.988, 0.0] clearly classifying it as unacceptable It means that the change in two features were ineffective in changing the model's decision which LIME failed to view while explaining that data instance. Explanation for instance (a') from LIME changes as shown in Fig.8.

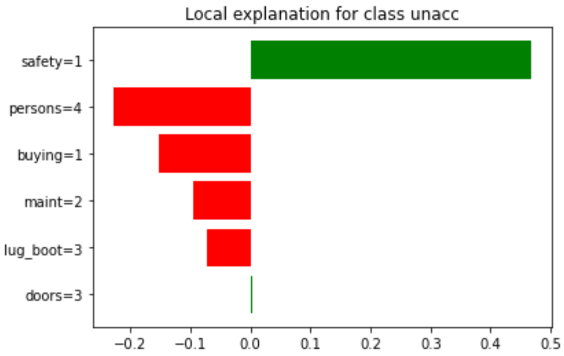


Fig.8. LIME explanation of instance (a')

But this non-influence of buying cost, luggage and boot space was already stated by the lattice in for instance (a) itself, proving that its explanation is absolutely correct. Explanations from the lattice consider the global perspective while explanations from LIME consider only the local neighborhood that do not provide a larger perspective.

Apart from being globally aware, lattice also generates similar and contrastive explanations that are quite intuitive to understand. For the instance (a) (*vhigh*, *med*, *3*, *4*, *small*, *low*), the similar and contrastive explanations are as follows:

Changing features: safety (low to high) changes the class to acc.

Changing features: buying (vhigh to high) does not change the class.

Changing features: buying (vhigh to low) does not change the class.

Changing features: buying (vhigh to med) does not change the class.

Changing features: maint (med to high) does not change the class.

Changing features: maint (med to low) does not change the class.

Changing features: doors (3 to 2) does not change the class.

Changing features: persons (4 to 2) does not change the class.

Changing features: lug_boot (small to big) does not change the class.

Changing features: lug_boot (small to med) does not change the class.

Go deeper?(y/n): y

Changing features: lug_boot (small to big) safety (low to high) changes the class to acc.

Changing features: lug_boot (small to med) safety (low to high) changes the class to acc.

Changing features: buying (vhigh to high) lug_boot (small to big) does not change the class.

Changing features: buying (vhigh to high) lug_boot (small to med) does not change the class.

Changing features: buying (vhigh to low) lug_boot (small to big) does not change the class.

Changing features: buying (vhigh to low) lug_boot (small to med) does not change the class.

Changing features: buying (vhigh to med) lug_boot (small to big) does not change the class.

Changing features: buying (vhigh to med) lug_boot (small to med) does not change the class.

Each of the above changes are easily verifiable by changing the feature value and checking the probability output of the Random Forest model. To test one such combination, "*Changing features: lug_boot (small to big) safety (low to high) changes the class to acc*", the model output for features: *buying=vhigh, maint=med, doors=3, persons=4,* ***lug_boot=big***, ***safety=high***, is [0.988, 0.0, 0.01, 0.002], classifying it as acceptable, and as stated by the lattice explanation.

4.2 INSTANCE FROM ACCEPTABLE CLASS

Instance (b): *buying=low, maint=med, doors=4, persons=more, lug_boot=small, safety=med*

The Random Forest model prediction probabilities for this data instance are [0.946, 0.04, 0.014, 0.0], classifying it as acceptable

4.2.1 *Explanation from LIME:*

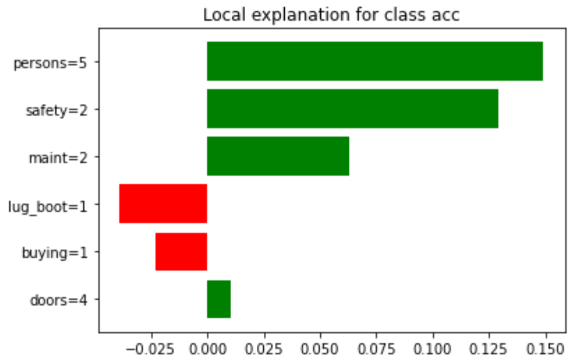


Fig.9. LIME explanation of instance (b)

In Fig.9, LIME identifies that since the car can carry five persons, with medium level of safety with low maintenance cost, it is acceptable with those three features playing the primary role, the number of doors being four also adds value. It also states that a small luggage and boot space has tried to pull it away from being acceptable While this part of the explanation is intuitive, LIME explanation also states that a low buying cost influences the car from not being acceptable, which appears counter intuitive. While having a small luggage and boot space may deem the car to be unacceptable, having a low buying cost may swing the car to a better class than acceptable and result in being good or very good. This counter intuitive part is not easy to comprehend from this explanation.

4.2.2 *Local Explanation from the Lattice:*

The following local explanation is obtained for instance (b) from the lattice:

Features: (lug_boot,small) deny class(es): vgood

Remaining class(es) are: acc good unacc

Features: (lug_boot,small) (safety,med) deny class(es): good

Remaining class(es) are: acc unacc

Features: (buying,low) (maint,med) (doors,4) (persons,more) (safety,med) deny class(es): unacc

Remaining class(es) are: acc

Lattice traversal has denied all class(es) except acc

Features: (persons,more) lead to class(es) acc with a confidence of 0.322917

Features: (persons,more) (safety,med) lead to class(es) acc with a confidence of 0.468750

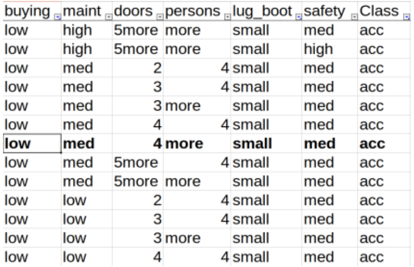
Features: (maint,med) (persons,more) (safety,med) lead to class(es) acc with a confidence of 0.604167

Features: (maint,med) (doors,4) (persons,more) (safety,med) lead to class(es) acc with a confidence of 0.666667

Features: (buying,low) (maint,med) (persons,more) (lug_boot,small) (safety,med) lead to class(es) acc with a confidence of 0.750000

Features: (buying,low) (maint,med) (doors,4) (persons,more) (lug_boot,small) (safety,med) lead to class(es) acc with a confidence of 1.000000

From the lattice explanation it is quite clear which feature combinations prevented the car from being very good, good or unacceptable A small luggage and boot space cannot make it very good, while a small luggage and boot space with medium safety cannot make it good.



buying	maint	doors	persons	lug_boot	safety	Class
low	high	5more	more	small	med	acc
low	high	5more	more	small	high	acc
low	med	2	4	small	med	acc
low	med	3	4	small	med	acc
low	med	3	more	small	med	acc
low	med	4	4	small	med	acc
low	**med**	**4**	**more**	**small**	**med**	**acc**
low	med	5more	4	small	med	acc
low	med	5more	more	small	med	acc
low	low	2	4	small	med	acc
low	low	3	4	small	med	acc
low	low	3	more	small	med	acc
low	low	4	4	small	med	acc

Fig.10. Dataset instances with buying=low, lug_boot=small and Class=acc

Since the acceptable class is the closest to the unacceptable class, there are a minimum of five feature combinations that deny it from being unacceptable In the part where features are listed to make it acceptable, the primary features listed are persons, safety, maintenance and doors which are exactly the same as the order of features with positive influence stated in the explanation from LIME. While LIME states that a low buying cost and a small luggage and boot space influence it against being acceptable, the last part of the explanation from lattice shows that these features in combination with the top four, influence it to be acceptable This disagreement is due to the fundamental assumption of independent features in LIME. But it is not so in real datasets and it is reflected rightly in the explanation from the lattice.

In order to check if this disagreement is an anomaly from the lattice and verify the truth, we filter the data with the criteria: *buying=low, lug_boot=small and Class=acc*.

According to LIME, these two features impact the decision of the model negatively from being acceptable But the filtered dataset in Fig.10 shows that there are multiple data instances that have this feature combination and are yet acceptable This proves that it is not an anomaly due to the lattice. It occurs as LIME generates random instances around the point of interest and weighs them proportionally based on its distance from the point of interest. Even if the randomly generated data produced the instances in Fig.10, from its explanation, we conclude that the weights of those were not large enough to modify its explanation.

But the explanation from lattice adds these features with the combination of other features to positively influence the decision of the model to be acceptable, which is evident from Fig.10 (bolded line, to be specific).

Apart from extracting feature combinations with their values, lattice also generates similar and contrastive explanations that are quite intuitive to understand. For the instance (b) (*low, med, 4, more, small, med*), the similar and contrastive explanations are as follows:

Generating similar & contrastive explanations:

Changing features: safety (med to high) changes the class to good.

Changing features: safety (med to low) changes the class to unacc.

Changing features: buying (low to high) changes the class to unacc.

Changing features: maint (med to high) does not change the class.

Changing features: maint (med to low) does not change the class.

Changing features: doors (4 to 2) changes the class to unacc.

Changing features: doors (4 to 3) does not change the class.

Changing features: persons (more to 2) changes the class to unacc.

Changing features: persons (more to 4) does not change the class.

Changing features: lug_boot (small to big) changes the class to good.

Changing features: lug_boot (small to med) changes the class to good.

Go deeper?(y/n): y

Changing features: lug_boot (small to big) safety (med to high) changes the class to vgood.

Changing features: lug_boot (small to big) safety (med to low) changes the class to unacc.

Changing features: lug_boot (small to med) safety (med to high) changes the class to vgood.

Changing features: lug_boot (small to med) safety (med to low) changes the class to unacc.

Each of the above changes are easily verifiable by changing the feature value and checking the probability output of the Random Forest model. Here, we verify the following three statements that claim a change of class:

Changing features: safety (med to high) changes the class to good.

Model output for (*buying=low, maint=med, doors=4, persons=more, lug_boot=small,* ***safety=high***) is [0.08, 0.884, 0.002, 0.034], classifying it as good, and as stated in this explanation.

Changing features: safety (med to low) changes the class to unacc.

Model output for (*buying=low, maint=med, doors=4, persons=more, lug_boot=small,* ***safety=low***) is [0.008, 0.004, 0.988, 0.0], classifying it as unacceptable, and as stated in this explanation.

Changing features: lug_boot (small to med) safety (med to high) changes the class to vgood.

Model output for (*buying=low, maint=med, doors=4, persons=more,* ***lug_boot=med***, ***safety=high***) is [0.006, 0.116, 0.0, 0.878], classifying it as very good, and as stated in this explanation.

4.3 INSTANCE FROM GOOD CLASS

Instance (c): *buying=low, maint=low, doors=2, persons=4, lug_boot=med, safety=high*

The Random Forest model prediction probabilities for this data instance are [0.036, 0.94, 0.004, 0.02], classifying it as good.

4.3.1 Explanation from LIME:

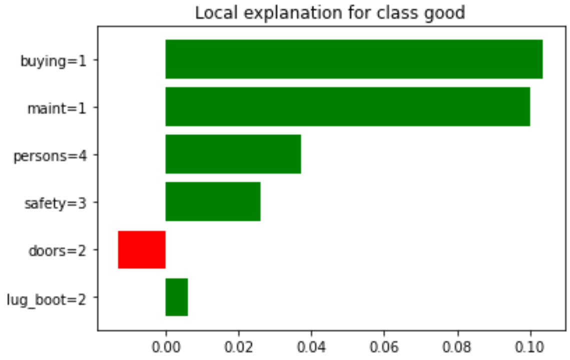


Fig.11. LIME explanation of instance (c)

In Fig.11, LIME identifies that since the car is low priced and can be maintained at low cost it is classified by the model to be good. Its capability to carry four persons and a high level of safety with a medium luggage and boot space also add value to be classified as good.

4.3.2 Local Explanation from the Lattice:

The following local explanation is obtained for instance (c) from the lattice:

Features: (doors,2) (lug_boot,med) deny class(es): vgood

Remaining class(es) are: acc good unacc

Features: (buying,low) (maint,low) (safety,high) deny class(es): acc

Remaining class(es) are: good unacc

Features: (buying,low) (persons,4) (safety,high) deny class(es): unacc

Remaining class(es) are: good

Lattice traversal has denied all class(es) except good

Features: (buying,low) lead to class(es) good with a confidence of 0.106481

Features: (buying,low) (maint,low) lead to class(es) good with a confidence of 0.212963

Features: (buying,low) (maint,low) (persons,4) lead to class(es) good with a confidence of 0.333333

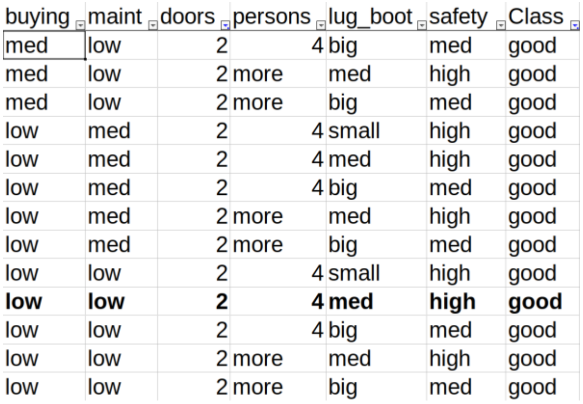
Features: (buying,low) (maint,low) (persons,4) (safety,high) lead to class(es) good with a confidence of 0.500000

Features: (buying,low) (maint,low) (doors,2) (persons,4) (safety,high) lead to class(es) good with a confidence of 0.666667

Features: (buying,low) (maint,low) (doors,2) (persons,4) (lug_boot,med) (safety,high) lead to class(es) good with a confidence of 1.000000

The lattice explanation lists primary four features of buying cost being low, maintenance cost being low, number of persons being four and high level of safety in that order with their combination to be classified by the model as good. This matches with the weights from the LIME explanation. Similar to the above cases, this instance where the number of doors are two, is seen with a negative influence by LIME as it considers the feature independent of others. But the lattice explanation states that in combination with other features stated above, it does lead the model to classify it as good.

It becomes clear when we filter the data with the criteria *doors=2, Class=good*



buying	maint	doors	persons	lug_boot	safety	Class
med	low	2	4	big	med	good
med	low	2	more	med	high	good
med	low	2	more	big	med	good
low	med	2	4	small	high	good
low	med	2	4	med	high	good
low	med	2	4	big	med	good
low	med	2	more	med	high	good
low	med	2	more	big	med	good
low	low	2	4	small	high	good
low	**low**	**2**	**4**	**med**	**high**	**good**
low	low	2	4	big	med	good
low	low	2	more	med	high	good
low	low	2	more	big	med	good

Fig.12. Dataset instances with doors=2 and Class=good

Fig.12 (bolded line, to be specific) shows that there are cars with two doors classified as good, but they are classified so in combination with specific values of other features.

Apart from extracting feature combinations with their values, lattice also generates similar and contrastive explanations that are quite intuitive to understand. For the instance (c) (*low, low, 2, 4, med, high*), the similar and contrastive explanations are as follows:

Generating similar & contrastive explanations:

Changing features: buying (low to high) changes the class to acc.

Changing features: maint (low to high) changes the class to acc.

Changing features: persons (4 to 2) changes the class to unacc.

Changing features: lug_boot (med to big) changes the class to vgood.

Go deeper?(y/n): y

Changing features: buying (low to high) lug_boot (med to big) changes the class to acc.

Changing features: maint (low to high) lug_boot (med to big) changes the class to vgood.

Changing features: persons (4 to 2) lug_boot (med to big) changes the class to unacc.

Changing features: buying (low to high) persons (4 to 2) changes the class to unacc.

Changing features: maint (low to high) persons (4 to 2) changes the class to unacc.

Changing features: buying (low to high) maint (low to high) changes the class to acc.

Each of the above changes are easily verifiable by changing the feature value and checking the probability output of the Random Forest model. Here, we verify one statement that claims a change of class to very good:

Changing features: lug_boot (med to big) changes the class to vgood.

Model output for (*buying=low, maint=low, doors=2, persons=4,* ***lug_boot=big****, safety=high*) is [0.02, 0.208, 0.004, 0.768], classifying it as very good, and as stated in this explanation.

4.4 INSTANCE FROM VERY GOOD CLASS

Instance (d): *buying=low, maint=low, doors=5more, persons=more, lug_boot=big, safety=high*

The Random Forest model prediction probabilities for this data instance are [0.006, 0.02, 0.002, 0.972], classifying it as very good.

4.4.1 Explanation from LIME:

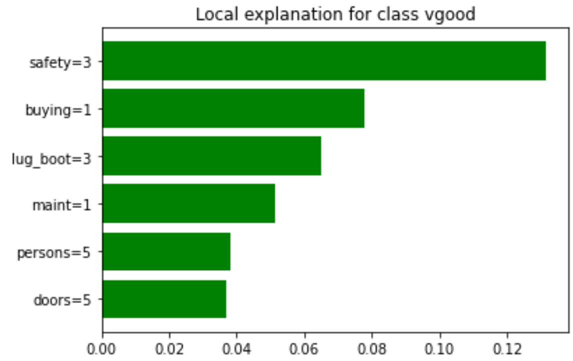


Fig.13. LIME explanation of instance (d)

In Fig.13, LIME explains that since all features are at their best, this car is classified as very good.

4.4.2 Local Explanation from the Lattice:

The following local explanation is obtained for instance (d) from the lattice:

Features: (lug_boot,big) (safety,high) deny class(es): good

Remaining class(es) are: acc unacc vgood

Features: (buying,low) (maint,low) (lug_boot,big) deny class(es): acc

Remaining class(es) are: unacc vgood

Features: (buying,low) (doors,5more) (persons,more) (safety,high) deny class(es): unacc

Remaining class(es) are: vgood

Lattice traversal has denied all class(es) except vgood

Features: (safety,high) lead to class(es) vgood with a confidence of 0.112847

Features: (buying,low) (safety,high) lead to class(es) vgood with a confidence of 0.270833

Features: (buying,low) (lug_boot,big) (safety,high) lead to class(es) vgood with a confidence of 0.500000

Features: (buying,low) (persons,more) (lug_boot,big) (safety,high) lead to class(es) vgood with a confidence of 0.750000

Features: (buying,low) (maint,low) (persons,more) (lug_boot,big) (safety,high) lead to class(es) vgood with a confidence of 1.000000

The lattice explanation matches with LIME for the first three features of high level of safety, low buying cost and a big luggage and boot space in that order. Beyond the first three features, while LIME lists low maintenance cost as the next feature in importance, the lattice considers the ability of the car to carry more persons. Low maintenance cost pictures after that in the lattice. Further LIME lists the number of doors being five at a weight quite close to the number of persons, while the lattice does not even consider this.

In order to check if this disagreement is an anomaly created by the lattice and to verify the truth, we check the ratio of instances that are classified as very good by adding the feature (*maint=low*) to the features (*safety=high, buying=low, lug_boot=big*) and to the total instances with that set of features and similarly by adding the feature (*persons=more*).

No. of instances with features (*safety=high, buying=low, lug_boot=big, maint=low*): 12

Out of these instances, no. of instances that classify as very good: 8

Confidence of this combination to be classified as very good is 8 / 12 = 66.67%

No. of instances with features (*safety=high, buying=low, lug_boot=big, persons=more*): 16

Out of these instances, no. of instances that classify as very good: 12

Confidence of this combination to be classified as very good is 12 / 16 = 75%

This calculation makes it clear that the feature (*persons=more*) added to the combination of (*safety=high, buying=low, lug_boot=big*) can influence the model to classify it as very good with a higher confidence than the feature (*maint=low*). This justifies the explanation from the lattice that lists the feature (*persons=more*) before (*maint=low*).

The second difference between the explanations from the lattice and LIME is that lattice does not even consider the feature (*doors=5*), while LIME gives it a weight close to (*persons=5*). In order to verify the truth behind this disagreement, we use the same features for the rest while altering the *doors* feature and check the probability output from the model.

Table.2. Model output for instances altering *doors* feature

buying	**maint**	**doors**	**per -sons**	**lug _boot**	**safety**	**Model Output**	**Class**
low	low	2	more	big	high	[0.008, 0.084, 0.008, 0.9]	vgood
low	low	3	more	big	high	[0.0, 0.038, 0.0, 0.962]	vgood
low	low	4	more	big	high	[0.006, 0.03, 0.0, 0.964]	vgood
low	low	5more	more	big	high	[0.006, 0.02, 0.002, 0.972]	vgood

Table.2 shows that the model classifies cars with any number of doors as very good in combination with specific values of other features. This is exactly stated in the explanation from the lattice which does not consider the doors feature in combination with other features.

Apart from extracting feature combinations with their values, lattice also generates similar and contrastive explanations that are quite intuitive to understand. For the instance (d) (*low, low, 5more, more, big, high*), the similar and contrastive explanations are as follows:

Generating similar & contrastive explanations:

Changing features: buying (low to high) changes the class to acc.

Changing features: maint (low to high) does not change the class.

Changing features: doors (5more to 2) does not change the class.

Changing features: doors (5more to 3) does not change the class.

Changing features: doors (5more to 4) does not change the class.

Changing features: persons (more to 2) changes the class to unacc.

Changing features: persons (more to 4) does not change the class.

Go deeper?(y/n): y

Changing features: buying (low to high) persons (more to 2) changes the class to unacc.

Changing features: buying (low to high) persons (more to 4) changes the class to acc.

Changing features: maint (low to high) persons (more to 2) changes the class to unacc.

Changing features: maint (low to high) persons (more to 4) does not change the class.

Changing features: doors (5more to 2) persons (more to 2) changes the class to unacc.

Each of the above changes are easily verifiable by changing the feature value and checking the probability output of the Random Forest model. Here, we verify one statement that claims a change of class to unacceptable

Changing features: persons (more to 2) changes the class to unacc.

Model output for (*buying=low, maint=low, doors=5more,* ***persons=2***, *lug_boot=big, safety=high*) is [0.0, 0.0, 0.988, 0.012], classifying it as unacceptable, and as stated in this explanation.

5. CONCLUSION AND FUTURE WORK

This comparative study clearly proves that the lattice-based approach to explanations is accurate and overcomes the limitations of LIME. A computationally efficient implementation of the lattice-based approach would be needed to compare it with LIME on larger datasets or to be applied on text and images. A comparative study with other existing state-of-the-art techniques can further prove its credibility and utility.

REFERENCES

[1] C. Rudin, “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead”, *Nature Machine Intelligence*, Vol. 1, No. 5, pp. 206-215, 2019.

[2] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila and Francisco Herrera, “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward responsible AI”, *Information Fusion*, Vol. 58, pp. 82-115, 2020.

[3] M.T. Ribeiro, S. Singh and C. Guestrin, “Why Should I Trust You? : Explaining the Predictions of Any Classifier”, *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016.

[4] Alvarez-Melis, David and Tommi S. Jaakkola, “On the Robustness of Interpretability Methods”, *Proceedings of International Conference on Machine Learning*, pp. 1-7, 2018.

[5] G. Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi and Davide Capuzzo, “Statistical Stability Indices for LIME: Obtaining Reliable Explanations for Machine Learning Models”, *Journal of the Operational Research Society*, Vol. 73, No. 1, pp. 91-101, 2022.

[6] Marzyeh Ghassemi, Luke Oakden Rayner and Andrew L Beam, “The False Hope of Current Approaches to Explainable Artificial Intelligence in Healthcare”, *The Lancet Digital Health*, Vol. 3, No. 11, pp. 745-750, 2021.

[7] S.M. Lundberg and S.I. Lee, “A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4765-4774, 2017.

[8] R.R. Selvaraju, A. Das and D. Batra, “Grad-CAM: Why did you say that?”, *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp. 354-356, 2016.

[9] D. Smilkov, N. Thorat, B. Kim and M. Wattenberg, “SmoothGrad: Removing Noise by Adding Noise”, *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp. 1-8, 2017.

[10] J.T. Springenberg and M.A. Riedmiller, “Striving for Simplicity: the all Convolutional Net”, *Proceedings of International Workshop on Information Communications*, pp. 1-6, 2015.

[11] M.L. Leavitt and A. Morcos, “Towards Falsifiable Interpretability Research”, *Proceedings of International Workshop on Neural Information Processing Systems*, pp. 98-104, 2020.

[12] M. Sundararajan, A, Taly and Q. Yan, “Axiomatic Attribution for Deep Networks”, *Proceedings of International Conference on Machine Learning*, pp. 3319-3328, 2017.

[13] J. Adebayo, J. Gilmer, M. Muelly and B. Kim, “Sanity Checks for Saliency Maps”, *Proceedings of International Conference on Neurocomputing*, pp. 9525-9536, 2018.

[14] Venkatsubramaniam Bhaskaran and Pallav Kumar Baruah, “A Novel Approach to Explainable AI Using Formal Concept Lattice”, *International Journal of Innovative Technology and Exploring Engineering*, Vol. 11, No. 7, pp. 1-17, 2022.

[15] A. Sangroya, M. Rastogi and L. Vig, “Guided-LIME: Structured Sampling based Hybrid Approach towards Explaining Blackbox Machine Learning Models”, *Proceedings of International Workshop on Computational Intelligence*, pp. 1-17, 2020.

[16] A. Sangroya, C. Anantaram, M. Rawat and M. Rastogi, “Using Formal Concept Analysis to Explain Black Box Deep Learning Classification Models”, *Proceedings of International Workshop on Machine Learning*, pp. 19-26, 2019.

[17] UCI, “UC Irvine Machine Learning Repository”, Available at: https://archive.ics.uci.edu/ml/index.php, Accessed at 2022.

[18] R. Wille, “Concept Lattices and Conceptual Knowledge Systems”, *Computers and Mathematics with Applications*, Vol. 23, pp. 493-515, 1992.

[19] UCI, “UCI Car Evaluation Data Set”, Available at: https://archive.ics.uci.edu/ml/datasets/Car+Evaluation, Accessed at 2022.