Interpretation of a Black Box Model with LIME and Clustering-LIME

Nikita Banga
Simon Fraser University
Burnaby, BC, Canada
nba45@sfu.ca

Animesh
Simon Fraser University
Burnaby, BC, Canada
aaa170@sfu.ca

Jian Pie
Simon Fraser University
Canada
jpie@sfu.ca

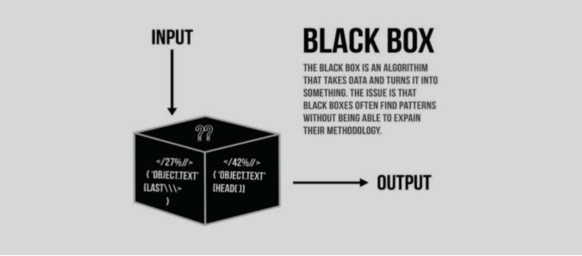
Abstract

In the era where technology is evolving and getting better day by day, we need to start questioning what we should choose to trust and what we shouldn't. Machine learning has become the topic of interest as a technical aspect in the industry. Although it seems appealing to let the machine handle the human world problems but how does it achieve its goals? Machine learning comprises of models that are complex in its working. These black box models are widely used but its working is still a matter of interest to researchers. There have been various algorithms that try to mimic the working of the black box models and make it interpretable to humans. LIME is one such algorithm that works in local areas of the data set and extracts the importance of features that are interlinked with the final output. In this paper we explore LIME and try to find another method to interpret black box models not only for a single data point but in the global aspect as well.

Keywords: Machine Learning, LIME, Local areas, black box models, interpretability

Introduction

Machine learning is a vast field that is associated with solving real work problems with the help of datasets. Initially, machine learning had simple models like naïve bayes and linear regression, which were mathematically interpretable. With time and exploration, machine learning became a complex field of study with the evolution of machine learning models. Machine learning was then divided into interpretable models and non-interpretable models (Black-box models). Black box model is a system whose complexity makes it difficult to understand the inward working of the model, for example, neural networks and gradient boost models. The Figure-1 below reflects the working of a black box model, where inputs and outputs are visible to the user but the inside working is not understandable. Such models made it difficult to understand how the it concluded with the predictions. Although the accuracy, precision and recall are useful measures to judge a model but it is not enough to explain the predictions. Let us suppose, we have a dataset which predicts the sleeping hours of an individual. Now we are able to fit a model with 98 percent accuracy but the question still remains that "How did it predict 7 hours of sleep for a person X?" "What factors affected the output?" or perhaps "Which factors played a major role in deciding the sleeping hours?". These questions are important when we want to know if the model is working as, it should or not. Blindly trusting a machine is not a responsible approach in solving a problem. What if the model is answering right because of some wrong features? Interpreting the model gives scope for improvement. For example, if we want to increase the sleeping hours of an individual and we know the factors that affected the person, we could easily find out how to increase the sleeping hours by working on those factors itself.



THE BLACK BOX IS AN ALGORITHIM THAT TAKES DATA AND TURNS IT INTO SOMETHING. THE ISSUE IS THAT BLACK BOXES OFTEN FIND PATTERNS WITHOUT BEING ABLE TO EXPAIN THEIR METHODOLOGY.

Figure 1 Black Box Models

LIME, Local Interpretable Model-Agnostic, was first coined in "Why Should I Trust You?" Explaining the Predictions of Any Classifier (KDD2016) by Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin [1]. The formal definition of LIME as provided by the paper is "It is a novel explanation technique that explains the prediction of a ML model in an interpretable and faithful manner by learning an interpretable model locally around the prediction" [1]. It became increasingly popular since it was able to interpret any black box model easily. It gave importance to the local aspect of the dataset to explain the global features. Our method is inspired by LIME to locally administer global black box model and explain its features, thereby, converting the complex model into a more interpretable model.

Motivation

Machine learning is growing exponentially and we are pushing our limits towards getting higher accuracy with so many complex models in the fields of creating intelligent machines. But what if a model is highly accurate but un-interpretable like Extreme gradient boost models which consist of ensembled tree structure which are highly complex and impossible to break down in smaller trees for getting a better understanding of its working, another example is a neural network which are highly accurate but at the same time consist of multiple densely connected layers of neurons that it is impossible to interpret the model by humans.

A very important question arises in this situation is that can we trust these models with high accuracy but very low or almost zero interpretability? this problem motivated us to work on one such model which is highly accurate and highly complex. There are already so many interpretation techniques available in the literature, multiple previous research papers have been written in attempt to interpret these complex models. We tried to take one such model which is local interpretable Model-agnostic explanation also called as LIME and tried to modify the basic principle behind the approach to create our own method to interpret complex models. We tried to find out the underlying working of Extreme Gradient boost models with the help of local linear regression models trained on the clusters of data points in same clusters individually and trying to interpret the working of XGBoost model on each individual chunk of datapoints. We also attempted to figure out the overall model feature importance by aggregating the results from each cluster.

LIME

LIME stands for local Interpretable Model-agnostic explanation, which is a model used to interpret any black box machine learning technique and it was introduced by Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. This technique appeared as a research paper in ACM's conference on Knowledge discovery and data mining in 2016. LIME explored precisely the question of model trust and interpretability.

Model-Agnostic is fancy term for saying that LIME can interpret any machine learning black box models. It attempts to understand the internal working of complex model by perturbing the input data samples and monitoring on how the model prediction changes. To understand LIME more precisely, let's take a black box model which takes input and gives a highly accurate output. LIME test's what will happen to the prediction when we give variation of our dataset into our model. It generates a new set of datapoints containing perturbed samples and the corresponding prediction of the black box model. LIME then trains a simpler model like linear regression or lasso,F there are so many simpler models available which they also call as a local model on this new dataset, and this local model is fairly simpler and easy to interpret. LIME uses the local model to find the important features for that specific prediction and gives the weightage to each feature along that point. This learned model may not be a very accurate interpretation of global model but it is highly reliable for local approximations.

LIME primarily focuses on two main aspects those are interpretable and local fidelity, interpretable means that the interpretation should provide a qualitative understanding between the input attributes and the model prediction. Local fidelity on the other hand deals with local interpretation of models, as there are no closed form solutions or explanations of these complex model and their working it is quite impossible for any model explanation to be completely exact description of model itself, so for an explanation to be acceptable it should at least have some meaningful explanation locally.

Local models mathematically can be expressed as follows:

$$\xi(x) = \operatorname{argmin} L(f, g, \pi x) + \Omega(g)$$

The explanation model for instance ‘x’ is the local model that minimizes the loss function, loss function basically calculates how far is the prediction is with the actual prediction of the global model (black box). ‘Ω’ denotes the model complexity which must be low for higher accuracy of interpretation. G is the family of possible explanations like all type of linear regression models. ‘π’ denotes how large we are taking the area around the chosen data point to generate new points, these generated data point also have a weight assigned to them according to the distance from the original datapoint in consideration.

Local Feature VS Global Feature

Machine learning model attempts to work with the whole dataset and interlinks the features with the outputs to make sense of the dataset to find patterns. This kind of work space is called global workspace that works with the whole dataset without considering categories within the dataset. When we try to categories the data and then formulate a model to find the outcome is called local workspace.

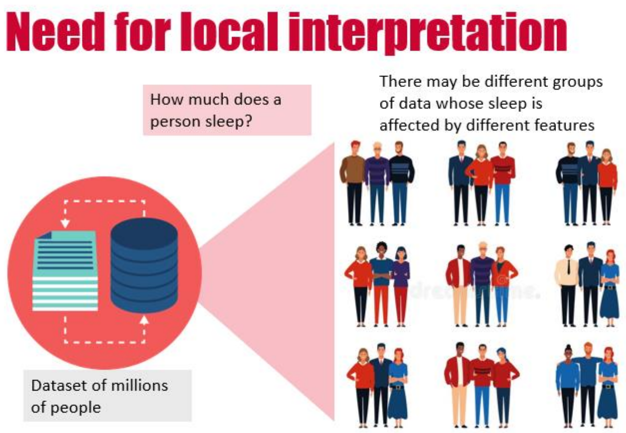

Need for local interpretation
How much does a person sleep?
There may be different groups of data whose sleep is affected by different features
Dataset of millions of people

Figure 2 Need for local Interpretation to understand the global Interpretation.

Before we move further, we need to understand why we should not interpret a big dataset directly. Let us assume we have a very big dataset and we want to know how much a person sleeps? Now in the whole data set, we have the accuracy, precision, recall and feature importance. But how do you know why a particular teenager sleeps less as compared to a baby? Those are different groups entirely. That means the dataset is full of data that has different kinds of people and its difficult to find the particular features impact on the outcome. In other words, the features for a teenager would be has a mobile, has school, plays games etc while for a baby it would be completely different, these features become irrelevant. Therefore, we need to break the problem into small groups of like data and work locally in finding the feature importance.

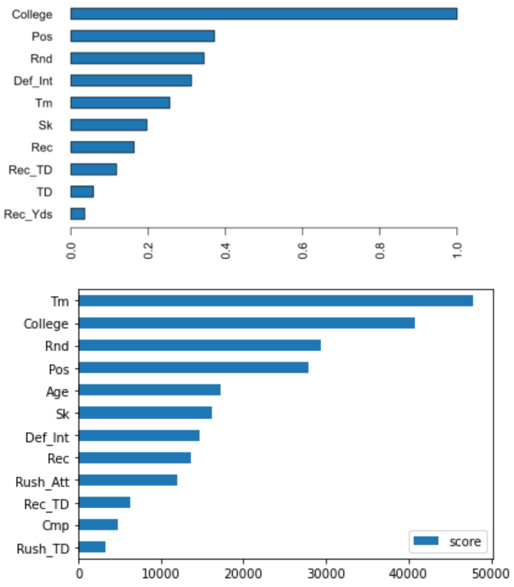


Figure 3 First figure shows the global importance of features while the second figure shows the local importance of features.

As it could be seen in the Figure 3 above, that global features are different from the local features. That explains the importance of categorising like-data , since one category might have different important features than other groups of data.

LIME, breaks the data set and makes sample instances around a datapoint to work locally and interpret a part of a very big dataset. As the name itself suggests, local interpretations are used to interpret the global model.

We will be working around the problem of making sample synthetic data around the data point by dividing the data before hand with the help of K-Medoids. After grouping like data together, we have localised the whole dataset and now it could be worked upon with linear interpretable models.

Experiment

We used two kinds of datasets:

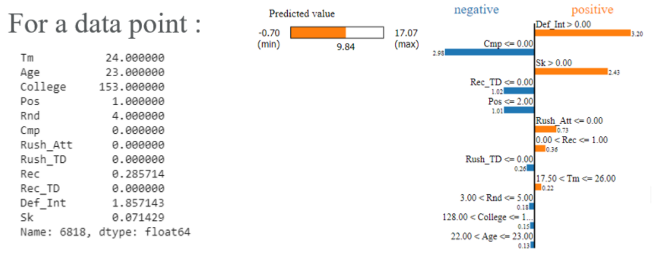
1. NFL: Taken from data.world with 20 features and a target being career length of the player. It comprises of approximately 8500 data points.
2. Used Bike prices: Taken from Kaggle with 6 features and a target being the price of the bike. Total data comprises of approximately 32000 data points.

For the conduction of experiments, we have used Google Collab combined with Jupyter notebook. The black box used is XGBoost that will be interpreted by LIME and our take on a new method.

Working with LIME

To understand the working of LIME, we experimented the two data sets and found out LIME interpretation of a datapoint belonging to the two datasets each.

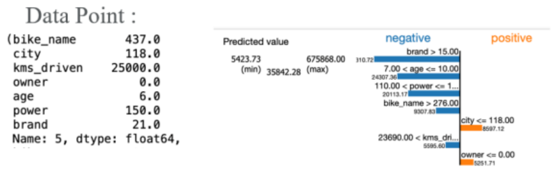
First we trained the model with XGBoost on NFL data to find out the career length of players. In the below Figure 4 we can see the LIME has given an interpretation of a single datapoint with the positive and negative impact of each feature and its correspondence to the output, that is, the career length for this particular feature. Since the age is between 22 to 23, completion is less than zero and position is less than 2, it causes the career to shorten in comparison to other features like team, sk,Def_Int that play a role in creasing the prediction. Overall, LIME has predicted approximately 10 years.


For a data point :

```
Tm            24.000000
Age           23.000000
College      153.000000
Pos            1.000000
Rnd            4.000000
Cmp            0.000000
Rush_Att       0.000000
Rush_TD        0.000000
Rec            0.285714
Rec_TD         0.000000
Def_Int        1.857143
Sk             0.071429
Name: 6818, dtype: float64
```

Figure 4 Lime interpretation with NFL Dataset

We then trained the same machine learning model, XGBoost , with Used Bike prices dataset. With the help of LIME we were able to understand the importance of features and its correlation to the prediction of prices for used bikes. In the below figure 5, we can see that the feature City and Owner are in the positive end while brand, KM driven and age are contributing to the negative side of the features graph affecting overall prediction of the model making it higher or lower respectively. Brand and age of the bike have a major negative impact on the price.



```
Data Point :
(bike_name      437.0
 city           118.0
 kms_driven   25000.0
 owner            0.0
 age              6.0
 power          150.0
 brand           21.0
 Name: 5, dtype: float64,
```

Figure 5 LIME interpretation of Used Car Dataset

Clustering-LIME

In LIME, a single data point is studied by making a sample test case and using a linear interpretable model to find the coefficients for features. We did not want to make test cases for each datapoint entered and wanted a more general representation of like data. Rather than making a synthetic dataset around a single data point as LIME does, we thought about dividing the dataset into clusters of like-data with K-Medoids. Clusters ensures that the like-data is together and has local features rather than global features. Here we calculated the clusters predictions with respect to XGBoost predictions for the same values. This step is important to select the optimal K value of the clusters. As seen below in Table 1, K = 4 gave the better average score over all other cluster approaches i.e. 3,5,6,7 therefore we

chose to divide our data in 4 clusters.

Table 1 Cluster R-Score for choosing the best number of clusters formation

K	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Average Score
4	0.59	0.42	0.53	0.49		0.507
5	0.17	0.51	0.49	0.31	0.33	0.360

One issue when building a linear model is correlated features. Correlated features can throw off importance measurements. We then perform a correlation analysis in each cluster. With the help of above correlation map we dropped highly correlated features.

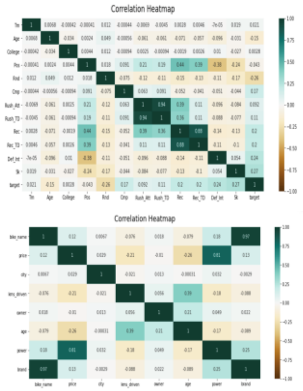


Figure 6 The correlation heat maps . Top figure represents Used Bike dataset while lower is for NFL Data.

The local model might not be in-sync with the global model. Therefore, it is good practice to compare the prediction of the linear interpretable model with the global predictions made on the same data points. We then worked on each cluster and compared its output with the global XGBooster prediction as given below in Figure 6.

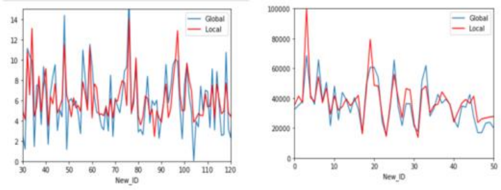
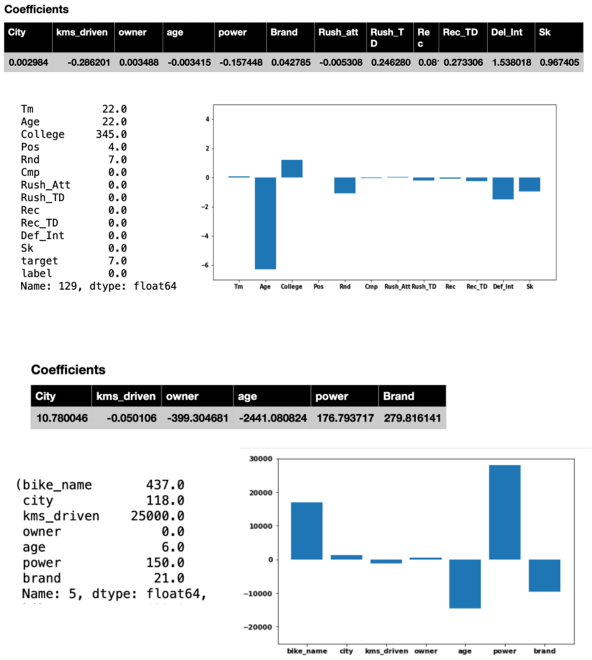


Figure 7 the comparison between local and global predictions. Left figure shows for NFL data and right figure is for Used bike dataset.

As seen in the Figure 7 The linear model is trying to fit the black box model pretty well. This means we can use this linear model to explain the global model. The reason they are not same is because they are linear models and not are not flexible when compare with complex model that can easily fit high dimensional and non-convex dataset as well.

We were able to find the coefficients for linear model and these coefficients are then used for interpreting a single datapoint within the cluster on the basis of the features impacting the output, as given in the figure 8 below.



Coefficients

City	kms_driven	owner	age	power	Brand	Rush_att	Rush_TD	Rec	Rec_TD	Del_Int	Sk
0.002984	-0.286201	0.003488	-0.003415	-0.157448	0.042785	-0.005308	0.246280	0.08	0.273306	1.538018	0.967405

```
Tm           22.0
Age          22.0
College     345.0
Pos           4.0
Rnd           7.0
Cmp           0.0
Rush_Att      0.0
Rush_TD       0.0
Rec           0.0
Rec_TD        0.0
Def_Int       0.0
Sk            0.0
target        7.0
label         0.0
Name: 129, dtype: float64
```

Coefficients

City	kms_driven	owner	age	power	Brand
10.780046	-0.050106	-399.304681	-2441.080824	176.793717	279.816141

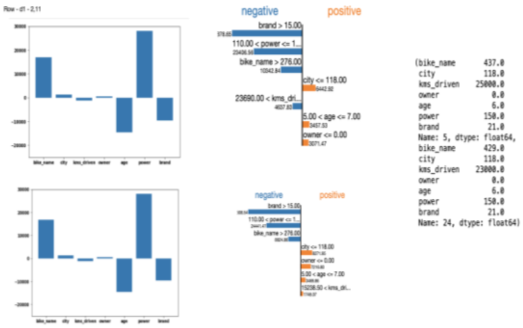
```
(bike_name       437.0
 city            118.0
 kms_driven    25000.0
 owner             0.0
 age               6.0
 power           150.0
 brand            21.0
 Name: 5, dtype: float64,
```

Figure 8 Interpretation of models after clustering and using linear regression.

Working with Cluster-LIME

1. Data points with similar attribute values

After performing LIME and Clustering-Lime with two similar points there were a lot of insight of the working of LIME. After using LIME, the two interpretations were completely different for the similar data points. If the points have similarity in the features, it is obvious that the same features should have the same effect on the outcome. But that was not the case after using LIME. We performed the same procedure for Clustering-LIME and it gave a better interpretation with little to no difference in similar features.



```
(bike_name       437.0
 city            118.0
 kms_driven    25000.0
 owner             0.0
 age               6.0
 power           150.0
 brand            21.0
 Name: 5, dtype: float64,
 bike_name       429.0
 city            118.0
 kms_driven    23000.0
 owner             0.0
 age               6.0
 power           150.0
 brand            21.0
 Name: 24, dtype: float64)
```

Figure 9 Interpretation analysis for experiment 1 for similar data points with LIME and Clustering-LIME.

2. Synthetically produced points with changes made to single feature.

For the second experiment, we took a data point and produced 2 similar data points by making changes in one feature. For experiment purpose, we change the age to 21,22 and 22.5. After performing the experiment, we saw drastic changes in LIME even though all the features were same except for the age. The age was also not changed with big numbers, that is, 21,22 and 22.5 are very close to each-other and should not affect the output much. The interpretation came out to be very different for each data point and the fun-fact being that the outcome was the same.

After applying Cluster-LIME, the observation was believable with little changes made to the age factor only, leaving others factors with no changes to the importance of factors, which seemed far more logical to understand the data points with similar data.

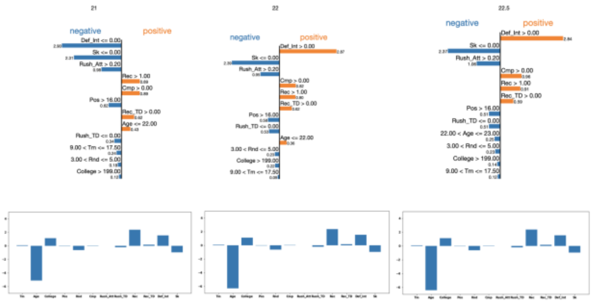


Figure 10 Interpretation analysis for experiment 2 for Synthetically produced points

3. Global features

XGBoost has a built- in method to find the features important for the prediction for the global set, but as we saw in figure 3**,** the global and local features might differ, therefore it is not very reliable. The second problem with the resultant feature importance chart is that there is no indication of the direction of the impact, the magnitude of the impact does not tell you much about the features as the direction of the feature importance might tell you.

We built a system to tell the magnitude and the direction (positive or negative) of the features of the dataset. Rather than going globally, we focused our attention to the different clusters and its coefficients. The coefficients of each cluster could be used to find the feature importance by taking the average. The figure 11 below shows the results of both the datasets. Although many features had a similar magnitude but it did not match with all the feature importance given by XGBoost.

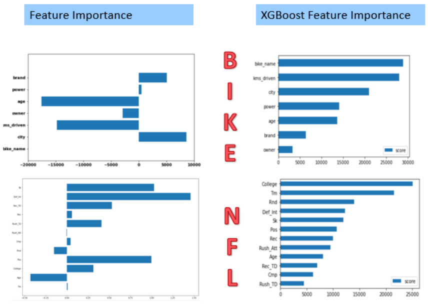


Figure 11 Cluster-LIME vs XGBoost Feature importance.

Conclusion

We compared our approach of clustering using K-Medoids technique with interpretation given by LIME. Lime basically works on a single data tuple and make data points around the tuple values with the help of perturbed datapoints and then using global model prediction it tries to fit a simpler model and find out the features and its impact on the output. Our approach of clustering the dataset does not create any new datapoints and works with original datapoints in the dataset. This approach is more reliable and also align with the prediction of global model output. LIME can also be a time consuming task especially if we have multiple datapoints to interpret and LIME need to go through the trouble of creating data points around each of the tuple and then fitting a local model into this new setting and interpret the feature importance while our approach already have cluster centers formed using similar datapoints among the cluster, we just need to find the nearest center for a particular datapoint and fit the model which is already trained for that specific cluster. It saves a lot of time and computation even if we have multiple points for interpretation. Using clustering approach, we also tried to find the global feature importance for black box models, not only the weight of how importance an attribute can be but also the direction in which it affects the model prediction. Global importance of black box model will give a general idea of which features are contributing positively on the final prediction of the black box models and which should impact the model accuracy negatively. Dataset are growing rapidly and figuring out the datapoint distribution is a complex task yet to be solved, so in case of dataset where datapoints do not have a cluster like distribution or all the points are sparsely distributed in the n dimensional space, K Medoids clustering approach will not be the most suitable model interpretation technique as it works on clustering phenomena. Global feature interpretation given from our approach works for many features comparing it to the inbuilt feature importance of model but does not work for all. This is quite a challenging task and it can be modified to get the better projection in future.

Future work

Cluster formation technique can be enhanced for making better cluster thereby increasing the interpretation accuracy for our approach can be a gamechanger. We tried using multiple linear regression models for finding the coefficients for each cluster, but we have not yet tested any decision tree models because it works way better for classification problems than regression intuitively. Using decision tree models as local models and finding a way for interpretation can be a future aspect for our approach.

References

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In KDD. ACM, 1135–1144.

[2] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpreting Blackbox Models via Model Extraction. arXiv:1705.08504 (2017).

[3]https://homes.cs.washington.edu/~marcotcm/blog/lime/

[4] Watson, D. (2020). Conceptual Challenges for Interpretable Machine Learning. *SSRN Electronic Journal*. Published. https://doi.org/10.2139/ssrn.3668444

[5] 林志.萍. (2021). An Improved Method of Interpret Machine Learning Based on LIME. *Hans Journal of Data Mining*, *11*(02), 38–49. https://doi.org/10.12677/hjdm.2021.112005