An Introduction on Interpretable Machine Learning

Neel Pradip Shah, Sheetal Jeshwani, Pavni Bhatt

Abstract*: As Artificial Intelligence penetrates all aspects of human life, more and more questions about ethical practices and fair uses arise, which has motivated the research community to look inside and develop methods to interpret these Artificial Intelligence/Machine Learning models. This concept of interpretability can not only help with the ethical questions but also can provide various insights into the working of these machine learning models, which will become crucial in trust-building and understanding how a model makes decisions. Furthermore, in many machine learning applications, the feature of interpretability is the primary value that they offer. However, in practice, many developers select models based on the accuracy score and disregarding the level of interpretability of that model, which can be chaotic as predictions by many high accuracy models are not easily explainable. In this paper, we introduce the concept of Machine Learning Model Interpretability, Interpretable Machine learning, and the methods used for interpretation and explanations.

***Keywords*: Machine Learning, Interpretability, Black Box Models, Explainable Artificial Intelligence**

I. INTRODUCTION

As more and more user-facing applications are using Machine learning and Artificial Intelligence, interpretability becomes more crucial than ever. One of the reasons is trust as most of these models are developed and trained by humans, so more the developers' trust their system, more the users will. Also, with interpretability, we, as users and developers, will understand and learn ways to use these models correctly, reducing model misuse. Apart from trust and fair use due to its universal adaptation, these machine learning models have caught the attention of various government agencies and policymakers. Many national governments have proposed and passed laws regulating the use of such models, such that the privacy of an ordinary citizen is maintained. Furthermore, many times these interpretations are used for meta-learning, like understanding general biases in the banking sector.

In this paper, we introduce the concept of Interpretable machine learning. We are starting with describing the concept of Interpretability from Machine learning or Artificial Intelligence perspective in section 2. In section 3, we will discuss the Interpretable models and the trade-off between model accuracy and interpretability. Then we will move towards the Model-Agnostic interpretation methods in section 4.

II. INTERPRETABILITY/EXPLAINABILITY

Interpretability means the degree to which a human can understand how a decision is made. A high interpretable Machine learning model means that the predictions made by that model are more comfortable to comprehend by humans. Many data scientists or engineers use the terms interpretable and explainable interchangeably. However, there is a clear distinction between them, and it is essential to understand this distinction before moving forward. As mentioned earlier, that interpretation means the understanding of how a decision is made. Furthermore, explanation means the understanding of why a decision is made [1].

A. Importance

Many times, people wonder why we need interpretation, why can't we accept the machine learning model and its outputs based on the accuracy scores. However, in many real-world situations, acceptance based on accuracy scores are not accepted, and that is a fact to digest. The application areas are such where we need meta-learning for generating more value, like in supermarkets for finding out frequently bought together, or in Ticket sales, for a particular event to analyze the buying patterns of the patrons, also, many times for satisfying human curiosity. Apart from that, interpretations play a major in detecting and analyzing model biases so that the engineers can tune the model for high-quality prediction.

B. Types of Interpretations

Global Interpretations

- ***Holistic Interpretability:*** This level of interpretability is about understanding how a model makes decisions by having a holistic view of the features, learned weights, and other parameters. For producing this level of interpretation, one must understand the algorithms and data, and the trained model itself.
- ***Modular-/Feature- Level Interpretability:*** From a big picture perspective, every feature plays a part in the prediction. Thus, understanding and interpreting the contribution of that feature towards the predictions can provide insights into how a feature or a set of features influence an output providing meta-information about the working of a model, and building trust.

Revised Manuscript Received on May 20, 2020.
***** Correspondence Author
Neel Pradip Shah*, Masters' Student, WIAI Faculty, University of Bamberg, Germany. E-mail: neel-pradip.shah@stud.uni-bamberg.de
Sheetal Jeshwani, Masters' Student, WIAI Faculty, University of Bamberg, Germany. E-mail: sheetal-jaikishan.jeshwani@stud.uni-bamberg.de
Pavni Bhatt, Data Analyst, MTLB India Private Ltd., India. E-mail: bhattpavni@gmail.com

Local Interpretations

- ***Single Prediction Interpretations***: In some applications areas, interpreting the model on a global level might be like using a sword to stitch a button, as these applications focus more on why a particular prediction? This kind of interpretation can help in understanding some complex constructs and feature relationships of the model better, as these interpretations depend only on a subset of the features.
- ***Interpretations for a group of predictions***: Let us say, for an image classifier, we want to understand why a particular set of images is put into a specific class? For such situations, one can use either global interpretations, considering that particular class of data as a whole dataset, or can interpret each of the decisions locally and combine them.

III. INTERPREATATBLE MODELS

In some applications areas, the simplest way to achieve interpretability is by using interpretable models. These machine learning models are a subset of machine learning models, which are humanly comprehensible. In applications like Bias free banking, Driverless Cars, Cancer detection use of such models is encouraged. In [1], the author provides an in-depth analysis of such models, and explain the association between features and target of these models the following table summarizes it,

Table- I: Interpretable Models and their Properties [1].

Models	Type	Linear	Monotone	Interaction
Linear Regression	Regression	Yes	Yes	No
Logistic Regression	Classification	No	Yes	No
Decision Trees	Classification, and Regression	No	To some extent	Yes
Rule Fit	Classification and Regression	Yes	No	Yes
Naive Bayes	Classification	No	Yes	No
K-Nearest Neighbour	Regression and Classification	No	No	No

In this section we will discuss in detail two such interpretable models Logistic Regression and Decision Trees.

A. Logistic Regression Model [1,2]

Logistic Regression is an extension of the linear regression model, which uses probabilities for classification problems with two possible outcomes. Linear Regression has one disadvantage when it comes to the classification that it considers the classes as numbers and tries to fit the best possible curve such that the distance between the points and the curve is minimal. Also, Linear models cannot work with probabilities.

Concept

Instead of fitting onto a straight-line logistic regression uses a log-based function to find the output of a linear function between 0 and 1,

$$\text{Logistic}(y) = \frac{1}{1+e^{y}}, \text{ where } y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \quad (1)$$

Visually this would look like,

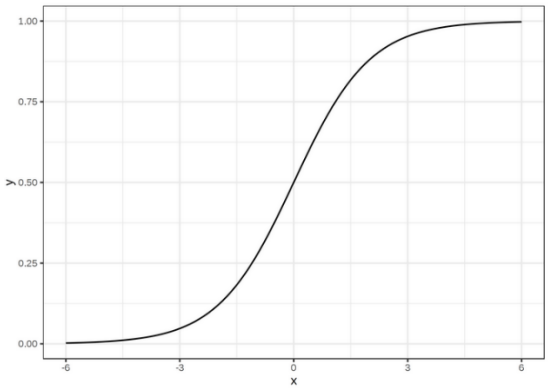

Fig. 1.Logistic Regression function

Interpretation

The primary difference between the interpretation of weights in linear and logistic regression is that in linear regression, the weights and features have a linear relationship, which is not the case in the logistic regression. Since the output of the logistic regression is a probability between 0 and 1, it is achieved by applying the logistic function to the linear weighted sum to transform it into probability values. Mathematically this can be described as follows,

$$log_e\left(\frac{P(y=1)}{1-P(y=1)}\right) = log_e\left(\frac{P(y=1)}{P(y=0)}\right) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \quad (2)$$

This can also be realized as,

$$\frac{P(y=1)}{1-P(y=1)} = e^{(a_0+a_1x_1+a_2x_2+\cdots+a_nx_n)} \quad (3)$$

After further study we can infer the following regarding to the interpretations for Logistic Regression

- Numerically if a feature value is changed by the magnitude of 1, the value of its respective weight changes exponentially
- For binary classification when a feature is classified from a class 0 to class 1, the odds is changed by $e^{\text{Weight of that feature}}$ factor.
- For base case i.e., when all the numerical values are set to 0 and the feature is classified for Class 0, the estimated odds are e^{a_0}.

Advantages and Disadvantages

Advantages,

- It Does not require too many computational resources.
- It is easy to implement and very efficient to train.

Disadvantages,

- Feature interactions are to be specified.
- Cannot solve non-linear problems.

- Identification all the important independent variables in advance is necessary.

B. Decision Tree Model [1,3]

Regression models fail in situations where the relationship between features and outcome is nonlinear and where features interact with each other [1]. In such cases, it better to use Decision trees. Decision trees recursively form smaller subsets of the dataset until a single set of outcomes is not reached based on a parameter. These subsets are nodes; basically, there are two kinds of nodes terminal (Correspond to a prediction) and leaf (Correspond to an output or an outcome) nodes.

Interpretation

In the case of the Decision tree interpretation is more human-friendly, one can understand the prediction, just by traversing the tree from the root node to the destined leaf node passing through the subsets and connecting each visited node with AND conjunction. While traversing the nodes, the edges provide information about the subset that is being traversed. The Feature importance in a decision tree can be computed using Information Gain (an Entropy-based method) and Gini Index. Also, one can trace individual prediction by doing a breakdown of each node visited during the traversal from the root node, as each of the nodes in the path has a certain amount of contribution to the final prediction.

Advantages and Disadvantages

Advantages,

- It is ideal for capturing interactions between features in the data.
- The tree structure also has a natural visualization, with its nodes and edges.
- Decision trees create good explanations.

Disadvantages,

- They are prone to overfitting.
- A few changes in the training dataset can create a completely different tree.
- Interpretability becomes an issue when the trees grow large.

C. Explainability vs Accuracy

Interpretability means to what extent humans can understand the prediction made by a model, and accuracy is the measure of correctness of the prediction made by a model. From a practical perspective, there is a trade-off between interpretability and accuracy, as many machine learning models enjoy a good accuracy score; at the same time, they are far more complex to be understood. Also, many models are interpretable they do not have a good accuracy score, as mentioned in [4] using a diagram like the following,

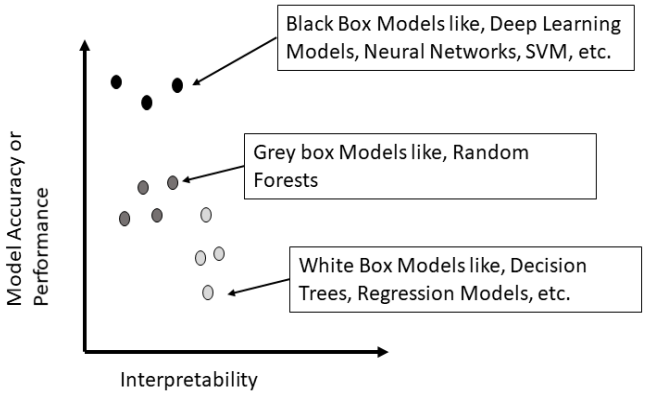

Fig. 2.Relationship between Interpretability and Accuracy

As described in the previous figure, namely high-performance models like Neural networks, Deep learning models (like CNN, RNN) performs better from an accuracy perspective. However, from the interpretability perspective, these models are complex to comprehend. Furthermore, models like decision trees and regression models are interpretable; these models lack in performance in comparison to black-box models. Also, the level of interpretability is not the same for all white-box models, as this trade-off can also be observed recursively in the individual set of these models.

IV. MODEL AGNOSTIC METHODS

Model-Agnostic Interpretation methods separate underlying machine learning models and the interpretation methods, and these methods provide flexibility in applying these methods to various machine learning models, using such methods provide not only a cost-effective way to have different permutations and combinations of various black-box models and interpretation methods but also flexibility in,

Representation sometimes features used to represent the instances are themselves not interpretable; in such cases, these model-agnostic methods can explain the output of the underlying black-box models using other features also.

Explanations, different kinds of explanations provide different information regarding the working of the model. As some explanations are more human-friendly than others, using such methods, we can try to explain our model to the maximum and not only make explanations more human-friendly but also can meet the regulations set by the authorities.

Model(s), White-Box, or Interpretable methods are not as accurate as Black-Box models. With the introduction of these interpretation methods, the developers will not be restricted to only Interpretable models to comply with regulations but will also be able to use Black-Box models providing good accuracy scores [1].

A. LIME [1, 4]

Proposed in 2016, LIME trains a surrogate model to estimate the predictions of the underlying Black-Box models for explaining individual predictions. Intuitively, LIME gives variations of the original data points as input to the model,

then observes the output and interprets them. Mathematically,

$$explanation(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (4)$$

Where,

$\mathbf{explanation}(x)$ is the explanation model

$\mathbf{argmin}$ is the point for which the function value is minimum

$\mathrm{L}(x)$ is the loss function which measures the distance between the predicted point and the interpretation.

f is the Blackbox model which we want to interpret

G is the set of possible interpretations

π_x is the proximity measure, which is basically distance between predicted values and the interpretation of a feature, calculated using distance matrix.

$\Omega(g)$ is the model complexity, means the number of features to consider.

Advantages and Disadvantages

Advantages,

- One can still use the same local model, even if the underlying model is replaced.
- One of the few methods which can work for Tabular, Text and images.
- More human friendly explanations, we can select the local model conducive to the user, by which the user can understand the black box by spending little time over it.

Disadvantages,

- Sampling is done disregarding the correlation between features, resulting into unlikely data points which can be used for training the local model.
- The neighbourhoods are very large, so to find the correct kernel width, brute force is to be used.
- Complexity of the local model is to be defined in advance.
- There has been an observation of instability in the explanations i.e., explanations of two very close data points can vary

B. SHAP [1, 5, 6]

Introduced in 2017, SHAP (SHapley Additive exPlanations) exploits the constructs of Shapley Values from Game Theory for determining the contribution of features in a prediction for generating explanations. Proposed by Shapley in 1958, **Shapley Values** [7] represent a fair distribution of payouts (from the total gains) to the participants of a coalition. Mathematically,

$$\phi_i(v) = \sum_{S \subset N-\{i\}} \frac{|S|!(N-|S|-1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right) \quad (5)$$

Where,

ϕ_i is the Shapley Value for a feature i

S is the subset of the features

N is the total number of features

$\mathbf{v}(x)$ is the cost/characteristic function, which provides prediction for feature values in set S that are marginalized over features that are not included in set S.

From a Machine Learning perspective, we can say that Shapley Values represent the average marginal contribution of a feature to the output. Which is what SHAP uses to determine the contribution of each feature for its global interpretations. Roughly SHAP Implementation carries out the following operations,

- Take a set of features(F) each of which can be attributed a value
- Calculate the permutations of F
- Calculate the marginal contribution given by a feature, this marginal Contribution is then used in determining the difference between the expected and original output data points of the Black-Box.

Sometimes this *marginal contribution* can be misinterpreted as the difference of the predicted values before and after the removal of that feature from model training, but it is the contribution of a feature to the difference between actual and means prediction. Which is what we can observe in the above equation (6). As the name suggests, SHAP represents the Shapley value explanations as an additive feature attribution, making SHAP as a Linear model.

Mathematically it can be represented as,

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \quad (6)$$

Where,

$\mathbf{g}(x)$ is the explanation model or local model

$\boldsymbol{\phi}$ is the shapely value

M is the Coalition Size

z' is the Coalition vector and $\mathbf{z'} \subset \mathrm{M}$

SHAP contains constructs for Local and Global Interpretations, for local interpretations, SHAP implementation has **KernelSHAP**, based on local surrogate models, it is a kernel-based approximation of Shapley Values, and for tree-based models, the implementation has **TreeSHAP**. Moreover, for global interpretations, SHAP contains methods based on the aggregation of Shapley Values. The SHAP implementation approximates the Shapley values as the computation of Shapley values is a time-intensive process.

Advantages and Disadvantages

Advantages,

- Efficient Implementation for tree-based models.
- Has Constructs for Global interpretations.
- Backed by strong theoretical concepts.
- Inherits features like, Fair Distribution, Efficiency, Symmetry additivity from Shapley Values.

Disadvantages,

- High Computing time.

- Risk of Misinterpretation
- Complexity of the local model is to be defined in advance.
- Feature Dependence is Ignored
- Access to Dataset is Needed

V. DISCUSSION

In order to achieve interpretability in machine learning models, there are two options; either use Interpretable models mentioned in section 3, who generate human-understandable predictions directly. Alternatively, in case of explaining black-box models, use model-agnostic interpretation methods like LIME and SHAP, which can explain, the predictions made by such black-box models. Moreover, as mentioned earlier also, in present times where machine learning is not used only for obtaining predictions, but for obtaining predictions which are understandable and agreeable, as the application area of the machine learning is expanding accommodating a lot of daily-life tasks, and that is where the trade-off between accuracy and interpretability comes into play, as the developers have to carefully select the models, keeping this trade-off and requirements in mind.

VI. FUTURE WORK

In this paper, we have introduced the concept of Interpretable machine learning and the methods by which we achieve interpretability in a machine learning model. Currently, this topic has gained momentum in the research community, as many national governments are proposing to regulate the use of data for training these models. Apart from the legal motivation, these techniques can provide information about the biases existing in the model. With this information, these models can build trust with the users and developers. After this performing this study, we will study the implementations of the model agnostic-methods provided by the authors of these methods to better understand their application by applying these methods on a trained model.

VII. CONCLUSION

In this paper, we have introduced the concept of Interpretable machine learning and its importance. Backed by these concepts, there are different methods and techniques are developed, and some existing models have been extended to accommodate interpretability. Through this paper, we have tried to provide a short yet rigorous introduction with the intent that it will provide a starting point to the community on this topic.

REFERENCES

1. Monlar Christoph, “Interpretable Machine Learning - A Guide for Making Black Box Models Explainable” Available: https://christophm.github.io/interpretable-ml-book/, 2019.
2. Kleinbaum David G., et al, Logistic regression. New York: Springer-Verlag, 2002.
3. Quinlan J. Ross, *Learning decision tree classifiers*. ACM Computing Surveys (CSUR) 28(1), pp.71-72, 1996.
4. Gunning David, Explainable Artificial Intelligence (XAI). Defence Advanced Research Projects Agency (DARPA)”, Available: https://www.darpa.mil/attachments/XAIProgramUpdate.pdf, 2017.
5. Ribeiro Marco T., Singh Sameer, Guestrin Carlos, Why should i trust you? Explaining the predictions of any classifier, In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp 1135-1144.
6. Lundberg Scott M., and Su-In Lee, A Unified Approach to Interpreting Model Predictions, arXiv preprint arXiv: 1705.07874, 2017.
7. Vojnović Milan, “Contest Theory - Incentive Mechanisms and Ranking Methods,” Cambridge University Press, 2016.