Interpretability in Machine Learning

Why we need to understand how our models make predictions

Conor O'Sullivan 1 day ago · 9 min read ★

Should we always trust a model that performs well? A model could reject your application for a mortgage or diagnose you with cancer. The consequences of these decisions are serious and, even if they are correct, we would expect an explanation. A human would be able to tell you that your income is too low for a mortgage or that a specific cluster of cells is likely malignant. A model that provided similar explanations would be more useful than one that just provided predictions.

Source: flaticon

By obtaining these explanations, we say we are interpreting a machine learning model. In the rest of this article, we'll explain in more detail what is meant by interpretability.

We'll then move on to the importance and benefits of being able to interpret your models. There are, however, still some downsides. We'll end off by discussing these and why, in some cases, you may prefer a less interpretable model.

What do we mean by interpretability?

In a previous article, I discuss the concept of model interpretability and how it relates to interpretable and explainable machine learning. To summarise, interpretability is the degree to which a model can be understood in human terms. Model A is more interpretable than model B if it is easier for a human to understand how model A makes predictions. For example, a Convolutional Neural Network is less interpretable than a Random Forest which is less interpretable than a Decision Tree.

With this in mind, we say a model is an interpretable model if it can be understood without any other aids/techniques. Interpretable models are highly interpretable. In comparison, explainable models are too complicated to be understood without the help of additional techniques. We say these models have low interpretability. We can see how these concepts are related in Figure 1. Generally, models can be classified as either interpretable or explainable but there are grey areas where people would disagree.

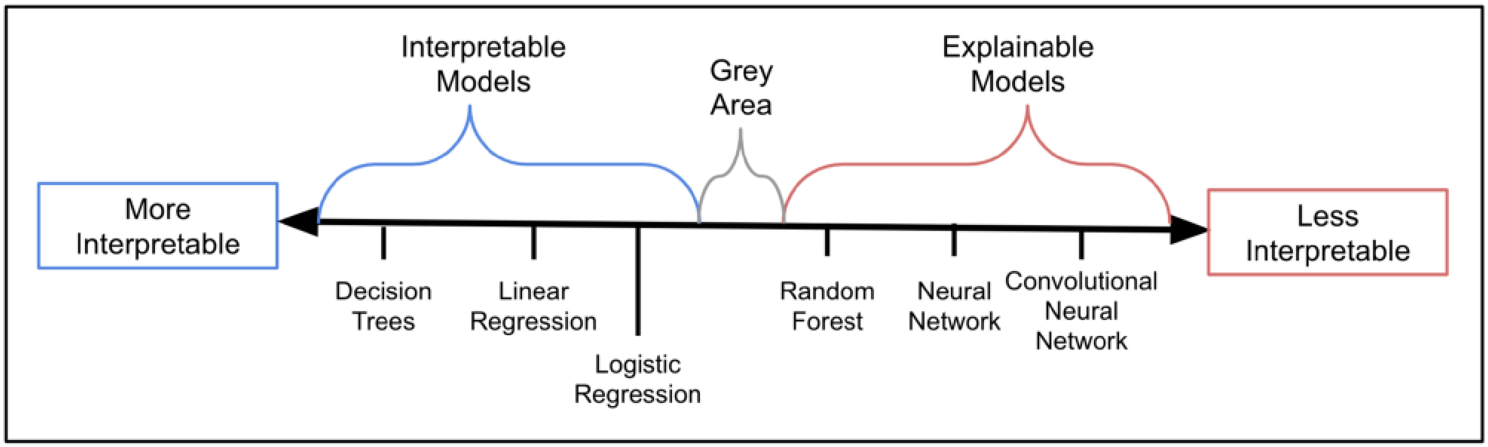
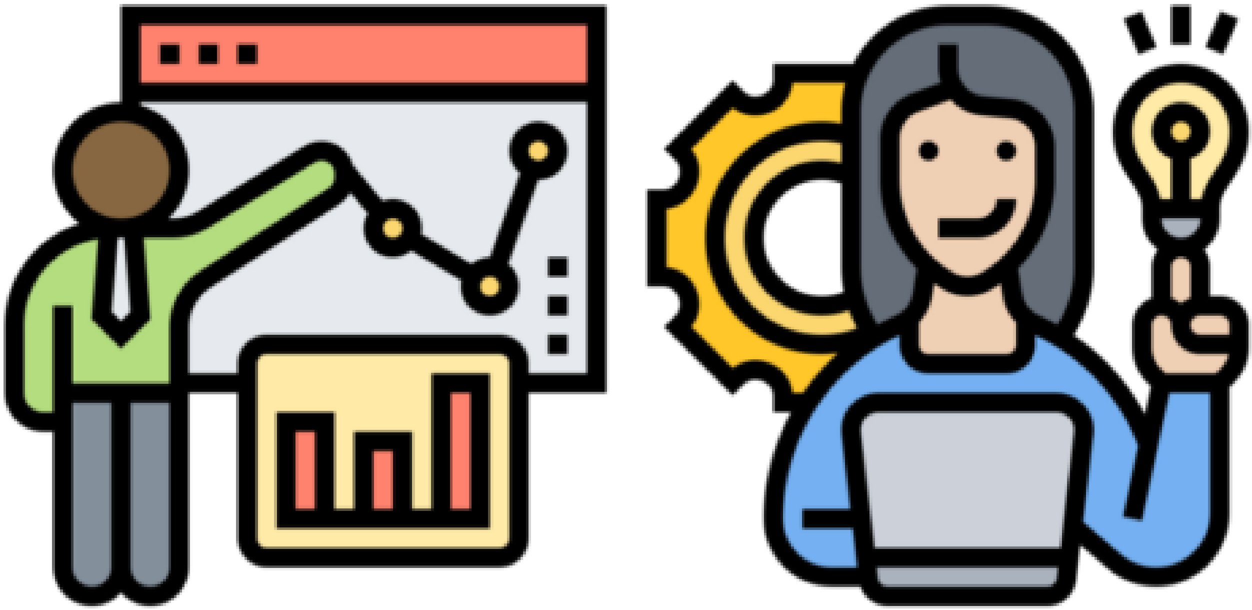


Figure 1: The Interpretability Spectrum (Source: Author)

Why is interpretability important?

As mentioned, we require additional techniques, such as feature importance or LIME, to understand how explainable models work. Implementing these techniques can be a lot of effort and, importantly, they only provide approximations for how a model works. So, we cannot be completely certain that we understand an explainable model. We can have a similar situation when comparing interpretable models.



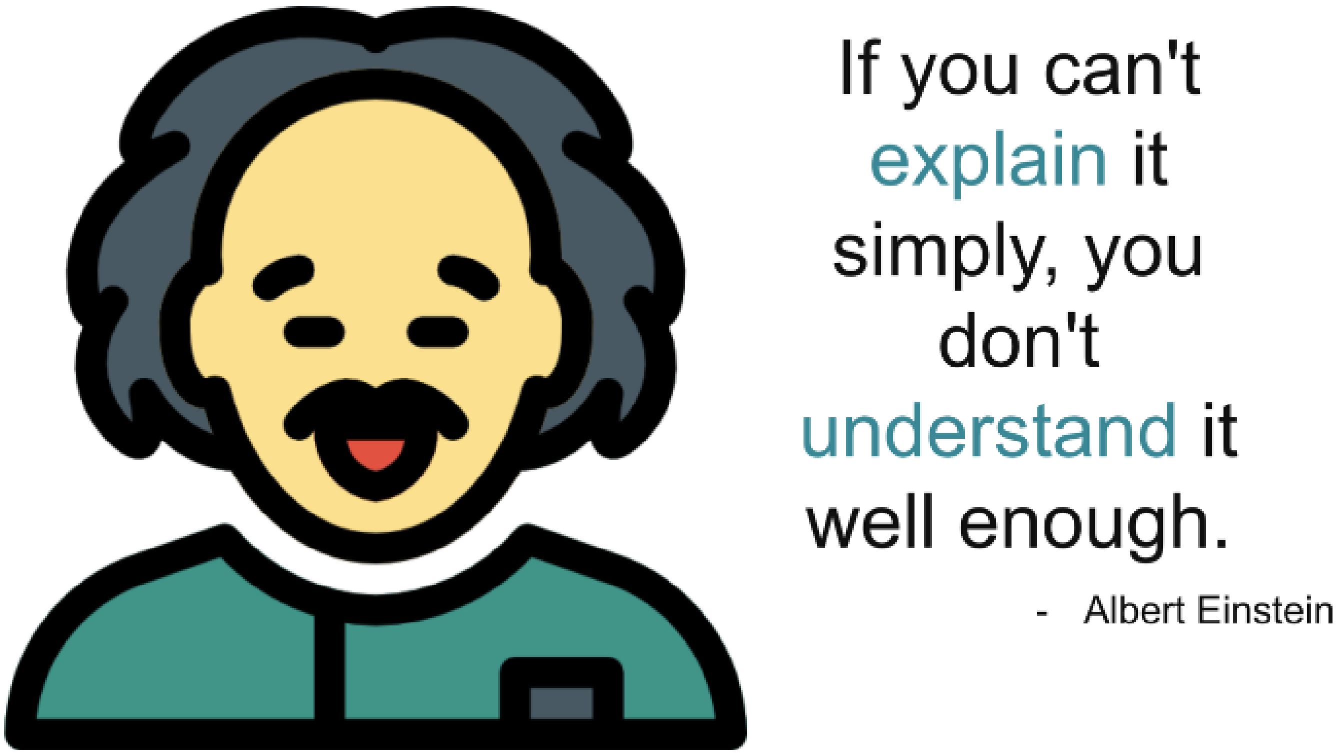
Source: flaticon

For example, logistic regression and decision trees. Neither of these requires additional techniques but logistic regression may still require more effort to interpret. We would need an understanding of the sigmoid function and how coefficients are related to odds/probability. This complexity may also lead to errors in our interpretations. In general, the more interpretable a model; the easier it is to understand and the more certain we can be that our understanding is correct. Interpretability is important because of the many benefits that flow from this.

Easier to explain

Our first benefit is that interpretable models are easier to explain to other people. For any topic, the better we understand it the easier it is to explain. We should also be able to explain it in simple terms (i.e. without mentioning the technical details). In industry, there are many people who may expect simple explanations for how your model works. These people will not necessarily have technical backgrounds or experience with machine learning.

For example, suppose we have created a model that predicts whether someone will make a life insurance claim. We want to use this model to automate life insurance underwriting at our company. To sign off on the model, our boss would require a detailed explanation of how it works. A disgruntled customer may rightly demand an explanation for why they were not approved for life cover. A regulator could even require such an explanation by law.

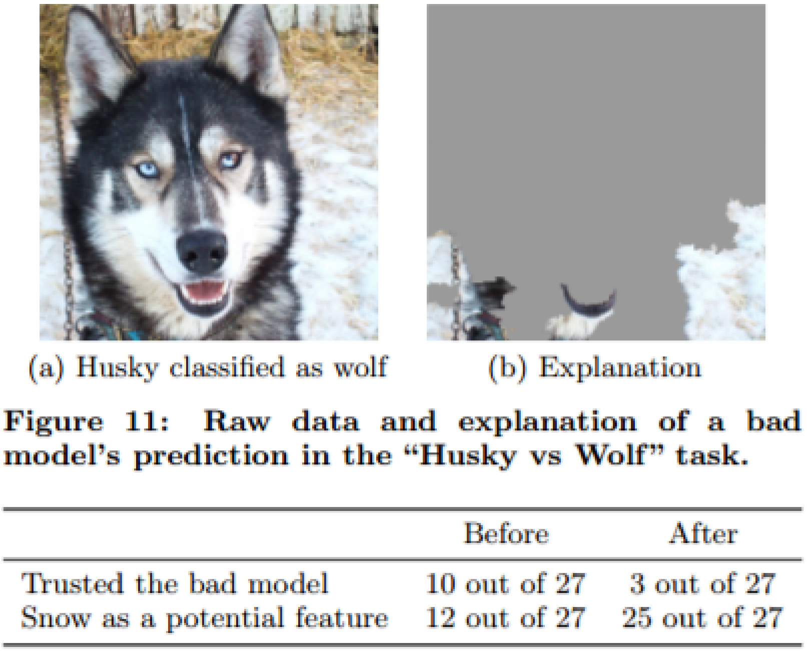

If you can't explain it simply, you don't understand it well enough.
- Albert Einstein

Source: flaticon

Trying to explain to these people how a neural network makes predictions may cause a lot of confusion. Due to the uncertainty, they may not even accept the explanation. In comparison, interpretable models like logistic regression can be understood in human terms. This means they can be explained in human terms. For example, we could explain precisely how much the customer's smoking habit has increased their probability of dying.

Easier to sense check and fix errors

The relationship described above is causal (i.e. smoking causes cancer/death). In general, machine learning models only care about associations. For example, a model could use someone's country of origin to predict if they had skin cancer. However, as with smoking, can we say someone's country causes cancer? The reason for this is that skin cancer is caused by sunshine and some countries are just sunnier than others. So we can only say skin cancer is associated with certain countries.



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Figure 2: Wolf vs husky experiment (Source: M. Tulio Ribeiro, S. Singh & C. Guestrin)

A good example of where associations can go wrong comes from an experiment performed by researches at the University of Washington. The researches trained an image recognition model to classify animals as a husky or wolf. Using LIME, they tried to understand how their model made predictions. In Figure 2, we can see that the model was basing its predictions on image backgrounds. If the background had snow, the animal was always classified as a wolf. They had essentially built a model that detects snow.

The issue is that wolves are associated with snow. Wolves will usually be found in the snow whereas huskies are not. What this example shows us is that models can, not only make incorrect predictions, but they can also make correct predictions in the wrong way. As data scientists, we need to sense check our models to make sure they are not making predictions in this way. The more interpretable your model the easier this is to do.

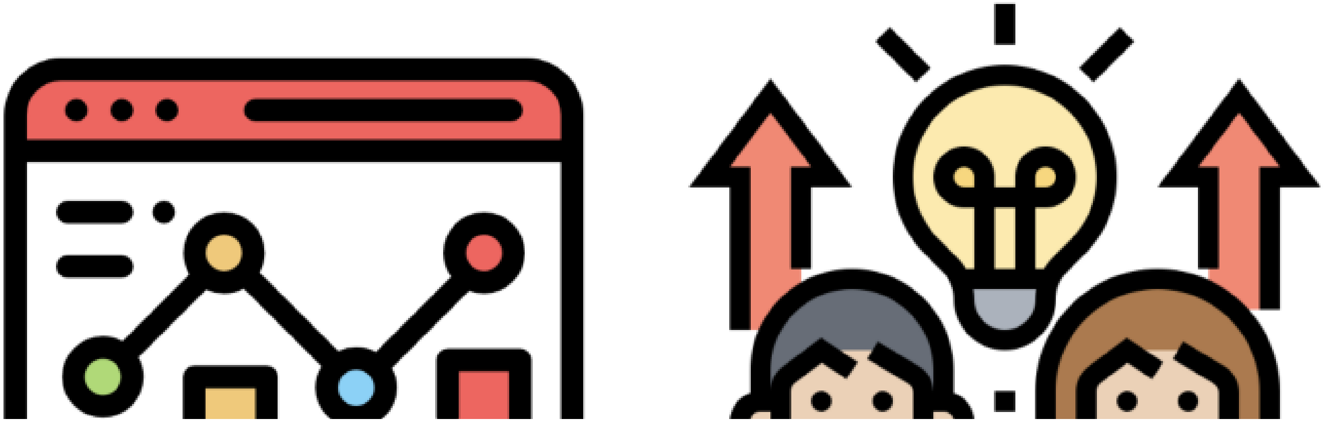
Source: flaticon

Easier to determine future performance

As time goes on, a model's prediction power may deteriorate. This is because relationships between model features and the target variable can change. For example, due to the wage gap, income may currently be a good predictor of gender. As society becomes more equal, income would lose its predictive strength. We need to be aware of these potential changes and their impact on our models. This is harder to do for explainable models. As it is less clear how features are being used, even if we know the impact on individual features, we may not be able to tell the impact on the model as a whole.

Easier to learn from the model

It is human nature to try to find meaning in the unknown. Machine learning can help us discover patterns in our data we didn't know existed. However, we cannot identify these patterns by just looking at the model's predictions. Any lessons are lost if we can not interpret our model. Ultimately, the less interpretable a model the harder it is to learn from it.

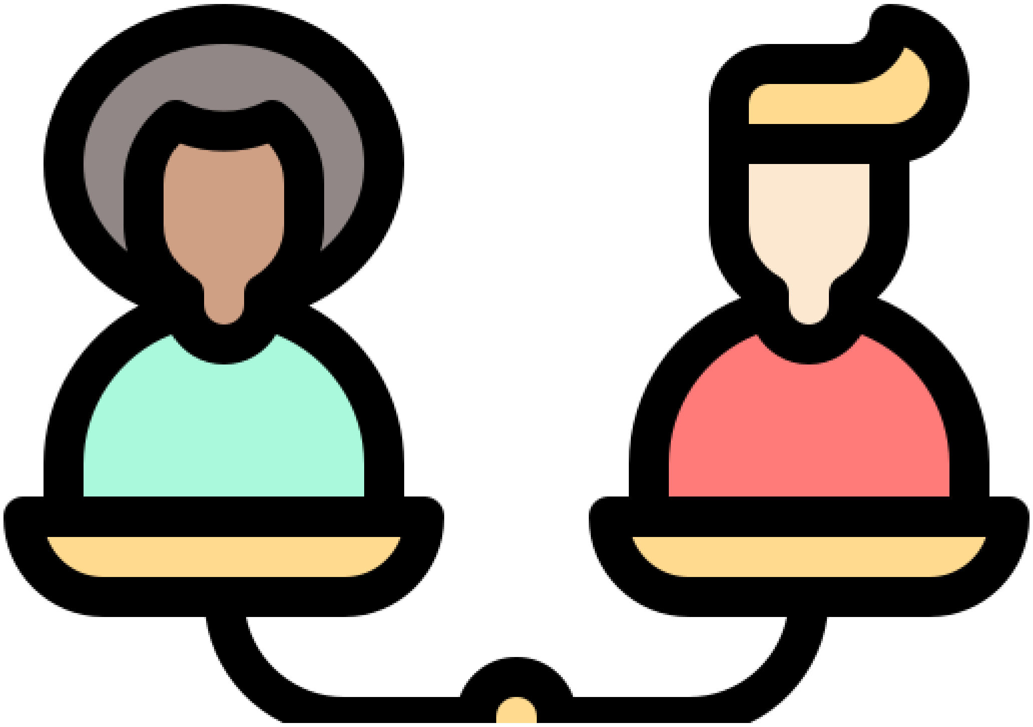


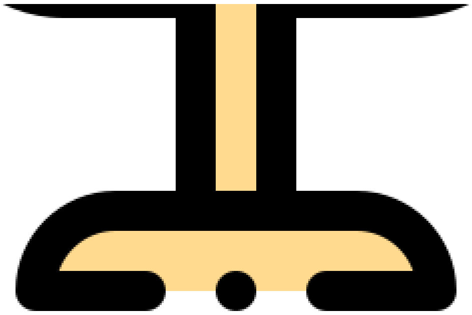
Source: flaticon

Algorithm Fairness

It is important that your models make unbiased decisions so that they do not perpetuate any historical injustices. Identifying sources of bias can be difficult. It often comes from associations between model features and protected variables (e.g. race or gender). For example, due to a history of forced segregation in South Africa, race is highly associated with someones' location\neighbourhood. Location can act as a proxy for race. A model that uses location may be biased towards a certain race.

Using an interpretable model will not necessarily mean that you will have an unbiased model. It also does not mean that it will be easier to determine if the model is fair or not. This is because most measures of fairness (e.g. false positive rate, disparate impact) are model agnostic. They are just as easy to calculate for any model. What using an interpretable model does do is make it easier to identify and correct the source of bias. We know what features are being used and we can check which of these are associated with the protected variables.





Source: flaticon

Downsides to interpretability

Okay, we get it… interpretable models are great. They are easier to understand, explain and learn from. They also allow us to better sense check current performance, future performance and model fairness. There are however downsides to interpretability and situations where we would prefer an explainable model.

Open to manipulation

Systems based on ML are open to manipulation or fraud. For example, suppose we have a system that automatically gives out car loans. An important feature could be the number of credit cards. The more cards a customer has the risker she is. If a customer knew this they could temporarily cancel all their cards, take out a car loan and then reapply for all the credit cards.

Source: flaticon

The probability of the customer repaying the loan does not change when she cancels her cards. The customer has manipulated the model to make an incorrect prediction. The more interpretable a model the more transparent and easy it is to manipulate. This is the case even if the inner working of a model are kept secret. The relationships between features and target variable are usually simpler making them easier to guess.

Less to learn

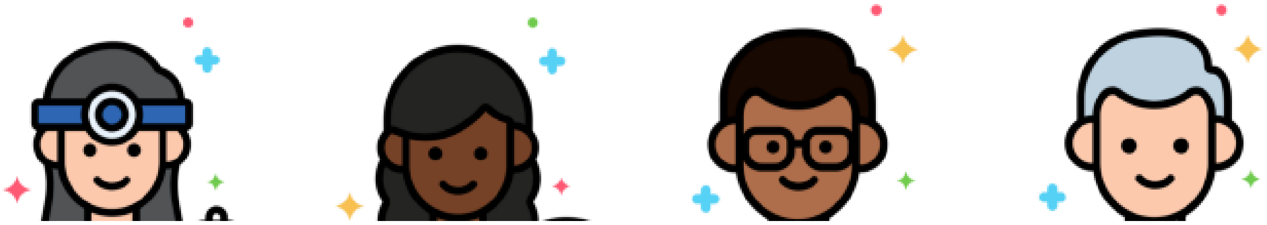
We mentioned that interpretable models are easier to learn from. The flip side is that they are less likely to teach us something new. An explainable model like a neural network can automatically model interactions and non-linear relationships in data. By interpreting these models we can uncover these relationships that we never knew existed.

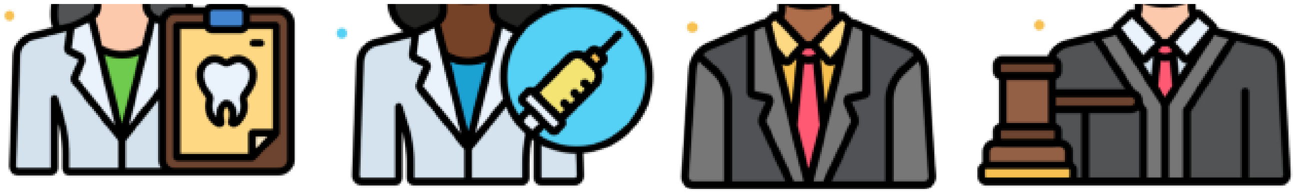
In comparison, algorithms like linear regression can only model linear relationships. To model non-linear relationships, we would have to use feature engineering to include any relevant variable in our dataset. This would require prior knowledge of the relationships defeating the purpose of interpreting the model.

Domain knowledge/ expertise requirement

Building interpretable models can require significant domain knowledge and expertise. Generally, interpretable models, like regression, can only model linear relationships in your data. To model non-linear relationships we have to perform feature engineering. For example, for a medical diagnosis model, we may want to calculate BMI using height and weight. Knowing what features will be predictive and, therefore, what features to create requires domain knowledge in a particular field.

Your team may not have this knowledge. Alternatively, you could use an explainable model which will automatically model non-linear relationships in your data. This removes the need to create any new features; essentially leaving the thinking up to the computer. The downside, as we've discussed thoroughly above, is a poorer understanding of how the features are being used to make predictions.





Complexity-Accuracy Trade-off

What we can see from the above is that, generally, the less complicated a model the more interpretable. So, for higher interpretability, there can be the trade-off of lower accuracy. This is because, in some cases, simpler models can make less accurate predictions. This really depends on the problem you are trying to solve. For instance, you would get poor results using logistic regression to do image recognition.

For many problems, an interpretable model would perform as well as an explainable model. In the article below, we compare an interpretable model, Logistic Regression, to an explainable model, a Neural Network. We show that by putting a bit of thought into our problem and creating new features we can achieve similar accuracy with an interpretable model. It is a good practical take on some of the concepts we've discussed in this article.

The Power of Feature Engineering

Why you should probably just use logistic regression to model non-linear decision boundaries (with Python code)

towardsdatascience.com

Image Sources

All images are my own or obtain from www.flaticon.com. In the case of the latter, I have a "Full license" as defined under their Premium Plan.

References

[1] C. Molnar, Interpretable Machine Learning (2020) https://christophm.github.io/interpretable-ml-book/

[2] C. Rudin, Stop explaining black-box machine learning models for high stakes decisions and use interpretable models instead (2019),

https://www.nature.com/articles/s42256-019-0048-x

[3] M. Tulio Ribeiro, S. Singh & C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016) https://arxiv.org/abs/1602.04938

Machine Learning Data Science Artificial Intelligence Interpretability Towards Data Science