Interpretability is crucial for trusting AI and machine learning

Ilknur Kaynar Kabul

10-12 minutes

We have updated our software for improved interpretability since this post was written. For the latest on this topic, read our new series on model-agnostic interpretability.

As machine learning takes its place in many recent advances in science and technology, the interpretability of machine learning models grows in importance.

We are surrounded with applications powered by machine learning, and we're personally affected by the decisions made by machines more and more every day. From the mundane to the lifesaving, we ask machine learning models for answers to questions like:

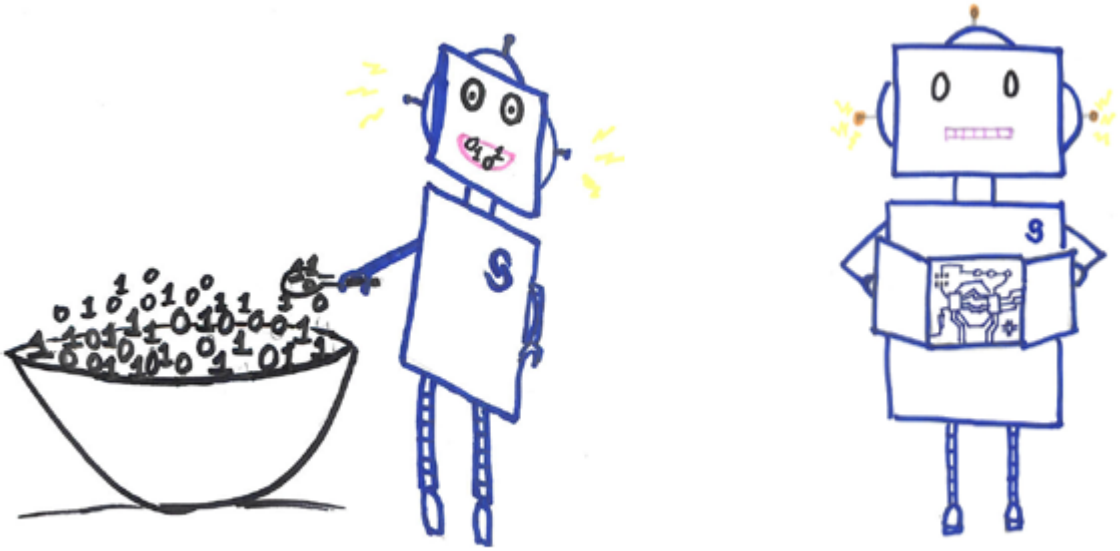
- What song will I enjoy?
- Will I be able to get a loan?
- Who should I hire?
- What is the probability of me having cancer?

These and many other questions are answered by predictive models that most users know very little about. Data scientists often put emphasis on the prediction accuracy of their models – not on understanding *how* those predictions are actually made. With machine learning, the models just do it.

Complex models are harder to understand

Some machine learning models are simple and easy to understand. We know how changing the inputs will affect the predicted outcome and can make justification for each prediction. However, with the recent advances in machine learning and artificial intelligence, models have become very complex, including complex deep neural networks and ensembles of different models. We refer to these complex models as black box models.

Unfortunately, the complexity that gives extraordinary predictive abilities to black box models also makes them very difficult to understand and trust. The algorithms inside the black box models do not expose their secrets. They don't, in general, provide a clear explanation of why they made a certain prediction. They just give us a probability, and they are opaque and hard to interpret. Sometimes there are thousands (even millions) of model parameters, there's no one-to-one relationship between input features and parameters, and often combinations of multiple models using many parameters affect the prediction. Some of them are also data-hungry. They need enormous amounts of data to achieve high accuracy. It's hard to figure out what they learned from those data sets and which of those data points have more influence on the outcome than the others.



Due to all those reasons, it’s very difficult to understand the process and the outcomes from those techniques. It’s also difficult to figure out whether we can trust the models and whether we can make fair decisions when using them.

What happens if they learn the wrong thing? What happens if they are not ready for deployment? There is a risk of misrepresentation, oversimplification or overfitting. Thus, we need to be careful when we use them, and we'd better understand how those models work.

Why accuracy is not enough

In machine learning, accuracy is measured by comparing the output of a machine learning model to the known actual values from the input data set.

A model can achieve high accuracy by memorizing the unimportant features or patterns in your data set. If there is a bias in your input data set, this can also affect your model. In addition, the data in the training environment may not be a good representation of the data in the production environment in which the model is deployed. Even if it is sufficiently representative initially, if we consider that the data in the production environment is not stationary, it can become outdated very quickly.

Thus, we cannot rely only on the prediction accuracy achieved for a specific data set. We need to know more. We need to demystify the black box machine learning models and improve transparency and interpretability to make them more trustworthy and reliable.

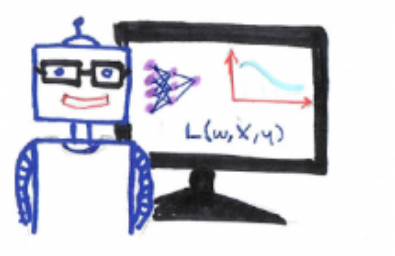
Interpretability means giving explanations to the end users for a particular decision or process. More specifically, it entails:

- Understanding the main tasks that affect the outcomes.
- Explaining the decisions that are made by an algorithm.
- Finding out the patterns/rules/features that are learned by an algorithm.
- Being critical about the results.
- Exploring the unknown unknowns for your algorithm.

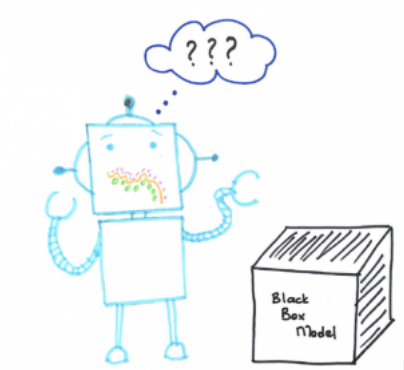
It is not about understanding every detail about how a model works for each data point in the training data.

Why do we need interpretability?

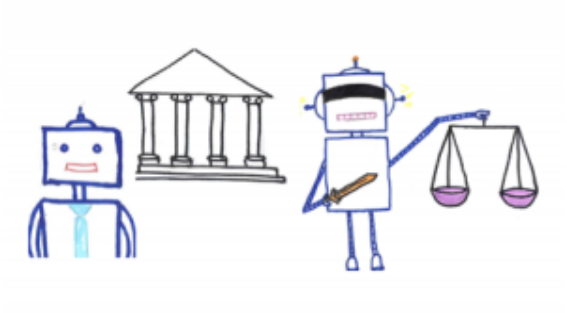
Interpretability is important to different people for different reasons:



Data scientists want to build models with high accuracy. They want to understand the details to find out how they can pick the best model and improve that model. They also want to get insights from the model so that they can communicate their findings to their target audience.



End users want to know why a model gives a certain prediction. They want to know how they will be affected by those decisions. They want to know whether they are being treated fairly and whether they need to object to any decision. They want to have a certain measure of trust when they are shopping online, or clicking ads on the web.



Regulators and lawmakers want to make the system fair and transparent. They want to protect consumers. With the inevitable rise of machine learning algorithms, they are becoming more concerned about the decisions made by models.

All those users want similar things from the black box models. They want them to be transparent, trustworthy and explainable.

1. **Transparent**: The system can explain how it works and/or why it gives certain predictions
2. **Trustworthy**: The system can handle different scenarios in the real world without continuous control.
3. **Explainable**: The system can convey useful information about its inner workings, for the patterns that it learns and for the results that it gives.

In a typical machine learning pipeline, we have control over the data set used to train the model, we have control over the model that we use, and we have control over how we assess and deploy those models.

When is interpretability needed?

If you need interpretability, first you need to ask yourself why you need it. In which stage of this process do you need interpretability? It may not be necessary to understand how a model makes its predictions for every application. However, you may need to know it if those predictions are used for high-stakes decisions. After you define your purpose, you should focus on what techniques you need in which stage of the process:

1. **Interpretability in pre-modeling (interpretability of model inputs):** Understanding your data set is very important before you start building models. You can use different exploratory data analysis and visualization techniques to have a better understanding of your data set. This can include summarizing the main characteristics of your data set, finding representative or critical points in your data set, and finding the relevant features from your data set. After you have an overall understanding of your data set, you need to think about which features you are going to use in modeling. If you want to explain the input-output relationship after you do modeling, you need to start with meaningful features. While highly engineered features (such as those obtained from t-sne, random projections, etc.) can boost the accuracy of your model, they will not be interpretable when you put the model to use.

2. **Interpretability in modeling:** We can categorize models as white box (transparent) and black box (opaque) models based on their simplicity, transparency and explainability.
    a. **White box (transparent) models**: Decision trees, rule-lists, and regression algorithms are usually considered in this category. These models are easy to understand when used with few predictors. They use interpretable transformations and give you more intuition about how things work, which helps you understand what's going on in the model. You can explain them to a technical audience. But of course, if you have hundreds of features and you build a very deep, large decision tree, things can still become complicated and uninterpretable.
    b. **Black box (opaque) models:** Deep neural networks, random forests, and gradient boosting machines can be considered in this category. They usually use many predictors and complex transformations. Some of them have many parameters. It's usually hard to visualize and understand what is going on inside these models. They're harder to communicate to a target audience. However, their prediction accuracy can be much better than other models. Recent research in this area hopes to make these models more transparent. Some of that research includes techniques that are part of the training process. Generating explanations in addition to the predictions is one way to improve transparency in these models. Another improvement is to include visualization of features after the training process.

3. **Interpretability in post-modeling (post hoc interpretability):** Interpretability in the model predictions helps us to inspect the dynamics between input features and output predictions. Some post-modeling activities are model-specific, while the others are model-agnostic. Adding interpretability at this phase can help us to understand the most important features for a model, how those features affect the predictions, how each feature contributes to the prediction and how sensitive your model is to certain features. There are model-agnostic techniques such as Partial Dependence (PD) plots, Individual Conditional Expectation (ICE) plots and Local Interpretable Model-Agnostic Explanations (LIME), in addition to the model-specific techniques, such as variable importance output from Random Forest.

Stay tuned for more posts on this topic. In this blog series, we'll go into more detail about some of the interpretability techniques. We'll explain how you can use model-agnostic interpretability techniques to understand black box models. We'll also cover some of the recent advances in interpretability.

Watch the webinar: Implementing AI Systems with Interpretability, transparency and trust