Q: Can we interpret a black-box model directly without knowing any information about the model itself?

A: It's a kind of model-agnostic explanation. If we need to describe the average behavior of the model without caring about the detailed individual predictions, one method can be employed is the global surrogate method, which trains an interpretable model to approximate the prediction of a black-box model. Its idea is that if an outcome of interest is difficult to measure directly, a cheap and interpretable surrogate model of the outcome can be used instead. Training a surrogate model is model-agnostic, and it consists of the following steps:

1. Select a dataset X. This can be the same dataset that was used for training the black-box model.
2. Get the predictions Y of the black-box model using the selected dataset X.
3. Collect a training dataset X' using the selected data X as input and the predicted outcomes Y from the black-box as the output.
4. Determine the interpretable model type like decision tree .
5. Train the interpretable model based on the collected training dataset X'.
6. Measure how well the surrogate model replicates the predictions of the black-box model, like R-squared measure.

After the surrogate model is trained, for new given input, interpret its corresponding output based on the interpretable model. The biggest advantage of the surrogate model method is that it's flexible which means various interpretable models can be employed. This is very helpful to provide model explanations to different people with different backgrounds and towards different application scenarios. Also, this method is quite straightforward, understandable, and easy to implement. However, since the surrogate models are only trained on the predictions of the black-box model instead of the real outcome, global surrogate models can only interpret the black-box model, but not the data. Also, determining what wellness should the surrogate model reaches is tricky and needs to be tried for distinct situations.

Reference

1. Molnar, Christoph. "Interpretable machine learning. A Guide for Making black-box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.
2. Molnar, Christoph, Giuseppe Casalicchio, and Bernd Bischl. "Interpretable machine learning–a brief history, state-of-the-art and challenges." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2020.