Q: If we are only interested in the explanation of a particular individual prediction, how to conduct such local interpreterbility without knowing the whole prediction model?

This is a kind of local interpretation problem. Local surrogate models can be employed as the interpretable models to explain individual predictions of black box machine learning models. It is to explain individual predictions of black-box models rather than the whole model. Local interpretable model-agnostic explanations (LIME) is a concrete implementation of local surrogate models, which tries to understand how the predictions change when the input data samples are perturbed. LIME generates a new dataset consisting of perturbed samples and the corresponding predictions of the black-box model. With this new dataset LIME then trains an interpretable model such as Lasso or decision tree, which is weighted by the proximity of the sampled instances to the instance of interest. The learned model should be a good approximation of the machine learning model predictions locally, but it does not have to be a good global approximation.
Likewise, the steps to train the LIME are as follows:

1. Select the instance of interest for which to have an explanation of its black-box prediction.
2. Perturb the dataset and get the black-box predictions for these new points.
3. Weight the new samples according to their proximity to the instance of interest.
4. Train a weighted, interpretable model on the dataset with the variations.
5. Explain the prediction by interpreting the local model.

How to get the variations of the data? It depends on the type of data, which can be either text, image or tabular data. For text and images, the solution is to turn single words or super-pixels on or off. In the case of tabular data, LIME creates new samples by perturbing each feature individually, drawing from a normal distribution with mean and standard deviation taken from the feature.
The advantage of local surrogate method includes the following aspects.
1) It's model-agnostic. You can still use the same local, interpretable model for explanation, even if you replace the underlying machine learning model.
2) Local surrogate models benefit from the literature and experience of training and interpreting interpretable models. When using Lasso or short trees, the resulting explanations are short and possibly contrastive. Therefore, they make human-friendly explanations.
3) The fidelity measure, that is how well the interpretable model approximates the black-box predictions, gives us a good idea of how reliable the interpretable model is in explaining the black-box predictions in the neighborhood of the data instance of interest.
As for the disadvantages:
1) The correct definition of the neighborhood is difficult when using tabular data. For each application, it has to try different kernel settings and see if the explanations make sense.
2) Current data points are sampled from a Gaussian distribution, ignoring the correlation between features. This can lead to unrealistic data points.

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.