Answer 9
Zhang Mingxue

In the machine learning world, there two kinds of models. One is called interpretable models. Another is called explainable models. The former could be understood by looking at its parameters, such as Linear Regression or Decision Tree. The latter is a much more complicated model containing many features or parameters, which people cannot understand and called a black box, such as a neural network [1]. However, in many situations, the prediction result of two kinds of models is more likely the same [2]. So, an argument that we do not need an explainable model instead should focus on interpreting models came out.

In Rudin's work [3], he proposed a new definition called the Rashomon set. The Rashomon set is the set of almost-equally-accurate models, which contains a large number of data. And as far as I am concerned, the Rashomon set is realistic. (Answered the first question)

Many high-risk decisions are not as complicated as possible, even though these complicated models may give us an explanation and seem to be correct. Although they have real meaning, many of these models are wrong, even their authors do not understand the basis for the decisions made by the model. [4] Thus, the scope of application of the models should be more subjective and low-risk. On the contrary, in high-risk decision-making, people should see the basis from the variables clearly. For example, it was mentioned in the paper [3] about whether to grant loans to people and bail decisions. If a pure black-box model is adopted, many people do not know why they are not allowed to loan because of the wrong input of their own procedures or logical conflicts. This is why explanatory models are not suitable for high-risk decisions. The interpretable models will tell us directly why we cannot get a loan like my age and occupation cannot be proven, so I can work to improve these aspects.

Many machine learning algorithms have many parameters to explain complex problems from a more complex perspective. Even though some features will support decision-making, they are still insignificant from a global perspective.[5] Therefore, an accurate model should be something we all know. The fundamental reason why there are many types of machine learning algorithms at present is that we have not fully understood them. For example, in the medical field, there are many current treatment methods for ankylosing spondylitis. Still, this illness has not been really overcome because medical scientists have not found the theoretical pathogenic factors. At present, they have only found the relevant reasons, but they cannot be truly determined to certain aspects. Therefore, in a sufficiently large machine learning model set, selecting a different number of parameters to predict a transaction, if the

Rashomon effect occurs, that is, the results obtained by each model are not much different, an optimal model can be found. So, the Rashomon set is realistic.

From the Rashomon set, we can get an accurate and interpretable model. When the most concise and precise model is found, other complex models will still get similar results due to the increase in the number of parameters or features, because when building predictive models of increasing complexity, the marginal gain from complicated models is typically small compared to the predictive power of the simple models. According to Hand's work, the simple models accounted for over 90% of the predictive power that could be achieved by "the best" model we could find. [6] Hence, to be accurate and interpretable, considering more parameters is unnecessary but increases the model's computational complexity. [7]

The theory obtained from this model can be reversed to obtain supplementary explanations for other models, which can be explained by concluding known variable types to other models. When there is no Rashomon effect between different models, that is, the models directly have their own explanation, which means that the current data is not extensive and comprehensive enough to produce truly correct results. In the computer vision applied in the automotive field, Tesla still has wrong results when judging the surrounding affairs because its training set is not complete enough to make the model get the most accurate answer. For example, when a car passes by a bicycle that objects obscure a large part, the algorithm cannot find similar objects in its training set, leading to its failure to identify. While a new method called ProtoPNet, which could make identification about birds and cars like humans, was proposed [8], its model is interpretable. The conflict between model results is still a matter of parameter selection and training set. Therefore, we can conclude that when there is a Rashomon set, it will definitely produce a situation where multiple models make similar conclusions. There must be the simplest model from which conclusions can be given to complex models to increase interpretability. In other words, the Rashomon set can be used to capture explainable models meaningfully. (Answered the second question).

Reference:

[1] Conor O'Sullivan.(2020) Interpretable vs Explainable Machine Learning. Web article at URL:
https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48

[2] Schmidt, L., Heße, F., Attinger, S., & Kumar, R. (2020). Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany. Water Resources Research, 56(5).
[3] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.
[4] Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research, 20(177), 1-81.
[5] Semenova, Lesia & Rudin, Cynthia. (2019). A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning.
[6] Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical science*, 1-14.
[7] Molnar, C. (2020). Interpretable machine learning. Lulu. com.
[8] Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., & Rudin, C. (2018). This looks like that: deep learning for interpretable image recognition. arXiv preprint arXiv:1806.10574.