All models are wrong, so which ones are useful?

*Przemysław BIECEK**

* MI2.AI, University of Warsaw and Warsaw University of Technology

George Box's quote "All models are wrong, some are useful," is a well-known phrase among statisticians. It acknowledges that it is impossible to create a perfectly accurate model to describe reality but that imperfect models can still be useful in real-world applications. Over the decades, various data-driven models have been developed using this philosophy and formulated to answer a variety of questions. Does the analyzed therapy produce positive medical outcomes? Does investment in education translate into student performance? These are just two examples of research questions that can be verified with data-driven models. Such approaches are the foundation of all empirical sciences.

G. Box, "Science and statistics", Journal of the American Statistical Association (1976)

Despite acknowledging the validity of Box's statement, an important question remains unanswered: how do we determine which models are useful? The answer is *some* but apparently not all of the models. Choosing the appropriate model or models is a critical decision. The conventional approach is to select some model quality criterion, usually based on how well the model fits the analyzed data, and then choose the model that best satisfies this criterion. There are several model quality criteria used by statisticians, including RMSE, R^2, AIC, BIC (we won't give their proper definitions here as they are of no importance to this article). In the machine learning community, criteria based on predictive performance on a new independent data are more common. However, the general procedure remains unaltered: start with a group of candidate models, select the best one according to a specific criterion and consider it the most accurate description of reality. From there we begin our inference.

F. Anscombe, "Graphs in Statistical Analysis", American Statistician (1973)

Such approaches sometimes lead to surprises and interesting paradoxes. One of them is *Anscombe's quartet*, introduced (precisely!) 50 years ago. Anscombe created four artificial sets of data, each consisting of 11 pairs of real numbers (like height and weight of eleven newborns). All those datasets have the same best-fitted linear model (in the sense of R^2) with the same value of R^2. Yet each set of data tells a completely different story. To understand the nature of the relationship between variables, visualization of the data is essential, as in Figure 1.

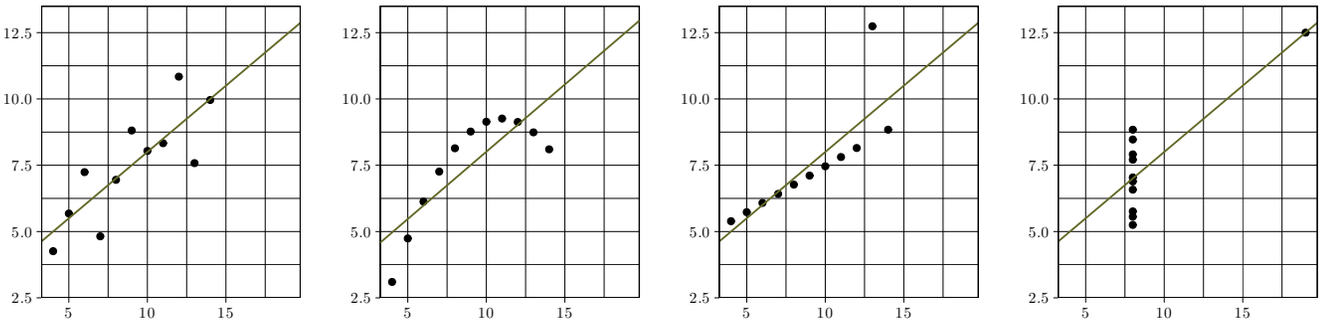


Fig. 1. Anscombe's quartet. For each dataset the model $y = x/2 + 3$ is the best linear fit and has similar fit to the data with the coefficient of determination $R^2 = 0.66$

As a consequence, a linear model that fits the data well is not enough to infer the relationship between variables accurately. Anscombe's solution to this paradox is to visualize the data, even with basic methods such as a scatterplot. Visual analysis can complement statistical inference in such cases and many well-known statisticians have proposed new methods of data visualization, which today are referred to as Exploratory Data Analysis tools.

J. Tukey, "Exploratory Data Analysis", Pearson (1977)

Anscombe demonstrated that different data sets can have the same model fit equally well but present entirely different stories. However, can the opposite be true? Can one dataset have several models with different stories that produce the same fit? Surprisingly the answer is positive. It was pointed out in 2001 by Leo Breiman in his influential paper "The Two Cultures". This quality is now

L. Breiman, "Statistical Modeling: The Two Cultures", Statistical Science (2001)

known as the "Rashomon perspectives" or "the multiplicity of good models" and it continues to exist in the foundations of statistical modeling in today's world, which is increasingly reliant on such models. The name "Rashomon" refers to a 1950 movie by Akira Kurosawa, in which an event is described from the perspective of four witnesses, each offering a different account of what had happened. Their stories vary so much that it is impossible to tell what is the truth. Breiman used this term to describe a hypothetical scenario in which several models have an equally good fit to the data, but they offer different explanations for the data. Such a situation would call into question any inference based on the "single best" data-driven model. For instance, what should one do facing two models that claim conflicting results about the effectiveness of a medical therapy? Which of these models should be trusted if they both have an equally good fit to the data?

In order to illustrate the multiplicity of good models problem, Breiman used several linear models with the same fit to the data (hardly different from the best possible linear fit). At the same time those model lead to different conclusions regarding the dependence between variables (e.g. is the increase of one variable followed by the increase or the decrease of another variable?) To make this phenomena even more striking, in our recent paper we introduced the *Rashomon's quartet*. The paper presents a regression tree, a random forest, a neural network, and a linear model (I will not delve into the details of these models here, for this article they are not important). All these models were fitted to the same data, resulting in the same predictive performance, but it turns out that each model is describing an entirely different story.

P. Biecek, H. Baniecki, M. Krzyziński, and D. Cook, "Performance is not enough: a story of the Rashomon's quartet" arxiv (2023)

... but wait! How do we know what stories are depicted by such complex models as a neural network or a random forest with hundreds of trees? Visualization techniques for predictive models developed under the name eXplainable Artificial Intelligence (XAI) or Explanatory Model Analysis (EMA) come to our aid. One of them is Partial Dependence (PD), a technique proposed by Jerome Friedman in his famous work on the boosting method. PD is a model agnostic method, meaning that it can be used to analyze any predictive model regardless of its complexity or structure. Due to this universality, this method has quickly found many applications.

J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", Annals of Statistics (2000)

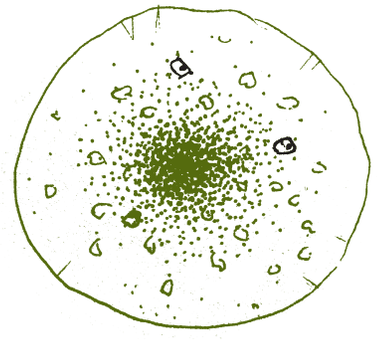
Let us introduce the intuition behind the Partial Dependence profile. The value of PD of any given variable S and its potential value t is equal to an average prediction of the model for the available data in which the value of the variable S is "artificially" set to t.

See also P. Biecek, T. Burzykowski "Explanatory Model Analysis" CRC Press (2021) or `https://ema.drwhy.ai/`

In order to describe it more formally let us assume that we have N observations of p input variables (e.g. for N newborns we observe weight, height, eye color...), where the i-th observation is $(x_{i,1}, x_{i,2}, \ldots, x_{i,p})$. These variables are used to predict *something* (this *something* is usually called the *target variable*) and in order to do so we apply the function $f(x_1, \ldots, x_p)$, which is our *model*. The Partial Dependence of the s-th variable is the function defined by the formula

$$PD^s(t) = \frac{1}{N}\sum_{i=1}^{N} f(x_{i,1}, ..., x_{i,s-1}, t, x_{i,s+1}, ..., x_{i,p}).$$

In essence, the PD shows how the predicted value of the target variable changes as the value of the predictor variable(s) of interest varies, while holding all other predictor variables constant at their observed values in the dataset.



The PD response profiles of each model in the Rashomon quartet are displayed in Figure 2, revealing distinct relationships between the variables, particularly for variables X_2 and X_3. For example, X_2 is insignificant in the first model but has positive effect in the other models. In contrast, X_3 is insignificant in the first model, has a negative effect in the second model and a positive effect in the third and fourth model. However, when all models have an equivalent fit to the data, it is challenging to determine which description is accurate.

Which model should we trust? If we don't know, why should we trust any of them? Since, due to random fluctuations, each of these model can be considered

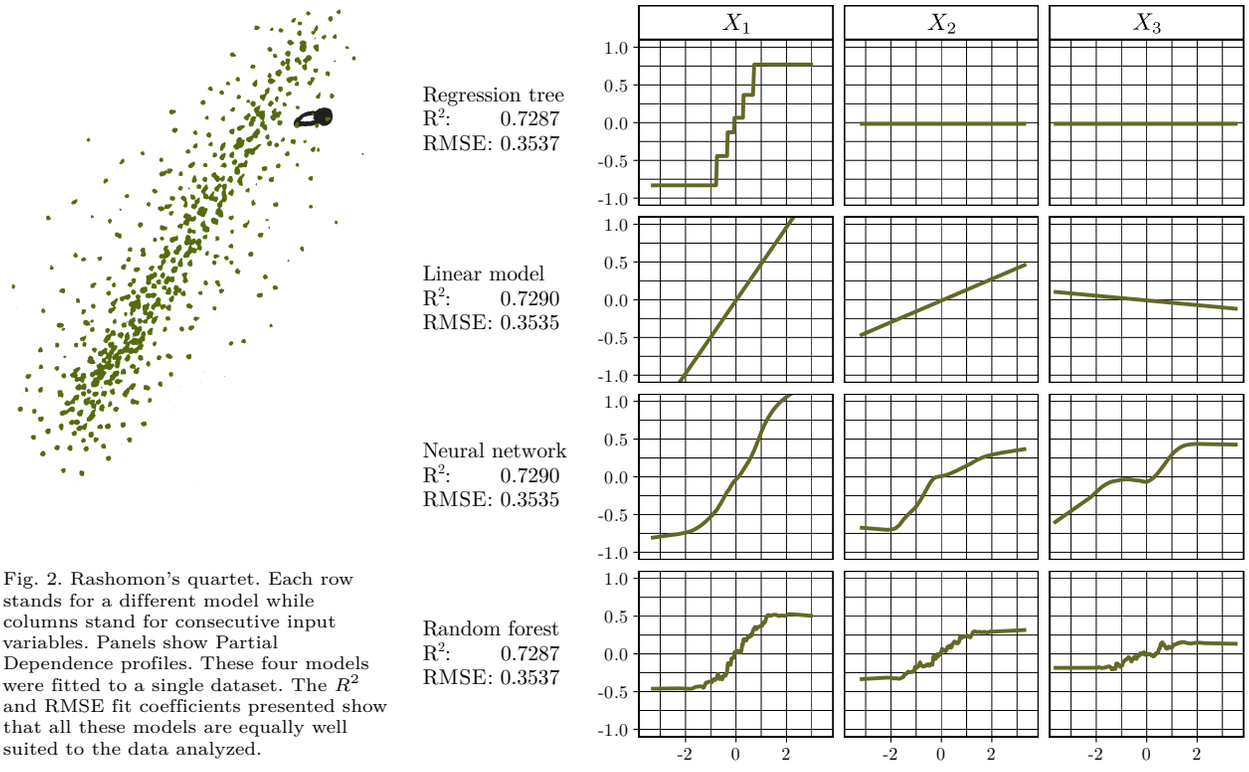


Fig. 2. Rashomon's quartet. Each row stands for a different model while columns stand for consecutive input variables. Panels show Partial Dependence profiles. These four models were fitted to a single dataset. The R^2 and RMSE fit coefficients presented show that all these models are equally well suited to the data analyzed.

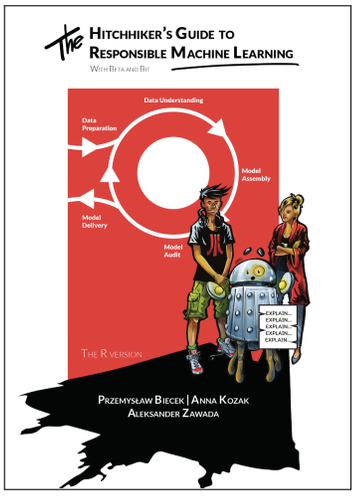
as the best one, analyzing only the single best model, and disregarding the slightly inferior ones, may be a bad strategy.

PD profiles say a lot about the models, but they are not yet the ultimate solution to the model visualisation problem. Despite their universality, they have some limitations or shortcomings, e.g. correlations between variables or interactions can distort the picture presented by these profiles. Variable-by-variable analysis will be difficult if the model uses hundreds or thousands of variables.

Characterization of a set of the best models is still an unresolved research problem, and different research groups are struggling with it. It is difficult but essential. To comprehend the world, we cannot depend solely on a single model's perspective, even if it has the best performance in relevant criteria. We must combine sets of models to differentiate between hypotheses supported by data and hypotheses that may result from the chosen predictive models.

C. Rudin, Ch. Chen, Z. Chen, H. Huang, L. Semenova, Ch. Zhong "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges" arxiv (2021)

The task of recognizing situations in which we can deal with the Rashomon perspective is yet to be solved. In the meantime, we can resort to thorough verification using a range of model visualization techniques. Anscombe's quartet has shown that data visualization techniques can be very useful for reasoning about nature of relations between variables. Similarly, Rashomon's quartet shows that model visualization techniques can be equally useful.



If you wish to explore the topic of model visualisation and comparison further, you may refer to resources such as the statistical comic book "The Hitchhiker's Guide to Responsible Machine Learning" available at `https://betaandbit.github.io/RML/`. Using the SARS-COV-2 mortality prediction as an example, it discusses the process of training predictive models as well as methods for exploring these models, including the best-known ones that is Partial Dependence, Shapley Values, and Ceteris Paribus. The examples presented above are for synthetic data. Similar challenges can be encountered when analyzing real-world problems as shown in the RML comic.

At the end of the day, we know that all models are wrong and we don't know which ones are useful. However, we can look at many good models at once. Looking at the world through the perspective of multiple models is essential to separate relationships truly supported by the data from relationships that are artifacts of the chosen modeling technique.