Введение

Одна из фундаментальных проблем машинного обучения в том, что обученные модели работают как "черный ящик". Машинное обучение в целом сводится к тому, что мы какую-то очень гибкую и настраиваемую функцию подгоняем под такой вид, который хорошо описывает какие-то имеющиеся данные. Например, мы знаем, как формируется нейронная сеть, знаем, как происходит её обучение, но после того, как она обучена, её поведение очень сложно интерпретировать - она может имеющиеся у нас данные описывать просто идеально, но мы всё равно не будем знать, какие именно принципы лежат в полученном функциональном преобразовании. Т.е. мы знаем выходной результат, но мы не знаем, почему мы получили именно такой результат, а не какой-то другой.

Задача - попытаться хоть как-то объяснить поведение таких моделей. Одна из сторон медали - попытаться для начала понять, какие из входных признаков являются наиболее значимыми по сравнению с другими. Есть модели, в которых эта задача решается "из коробки" просто из-за специфики построения модели - например, деревья принятия решений или случайные леса. Есть другие модели, типа нейронных сетей, в которых эта задача вообще никак не решается.