SHAP

SHapley Additive exPlanations

ERP: baopanpan3

姓名：包盼盼

部门：算法应用部

日期：2023年07月04日

目录

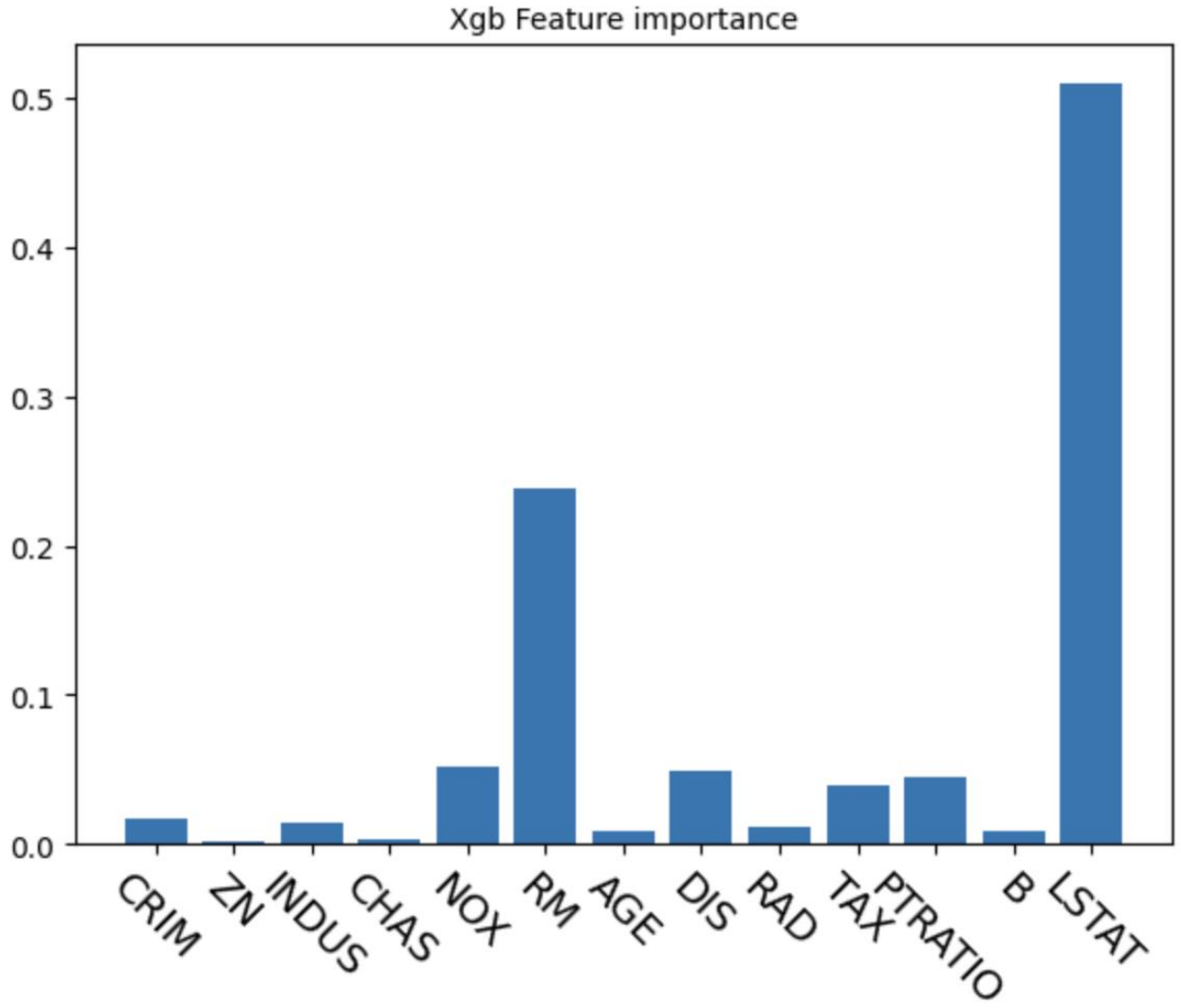
数据集-波士顿房价预测

CRIM - 城镇人均犯罪率
ZN - 占地面积超过25,000平方英尺的住宅用地比例
INDUS - 每个城镇非零售业务的比例。
CHAS - Charles River边界虚拟变量（如果是河道，则为1，否则为0）
NOX - 一氧化氮浓度
RM - 每间住宅的平均房间数
AGE - 1940年以前建造的自住单位比例
DIS - 到波士顿的五个就业中心加权距离
RAD - 径向高速公路的可达性指数
TAX - 每10,000美元的全额物业税率
PTRATIO - 城镇的学生与教师比例
B - 城镇黑人的比例
LSTAT - 房东属于低等收入阶层比例
MEDV - 自有住房的中位数报价(单位1000美元)

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33
...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1.0	273.0	21.0	391.99	9.67
502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1.0	273.0	21.0	396.90	9.08
503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1.0	273.0	21.0	396.90	5.64
504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1.0	273.0	21.0	393.45	6.48
505	0.04741	0.0	11.93	0.0	0.573	6.030	80.8	2.5050	1.0	273.0	21.0	396.90	7.88

506 rows × 13 columns

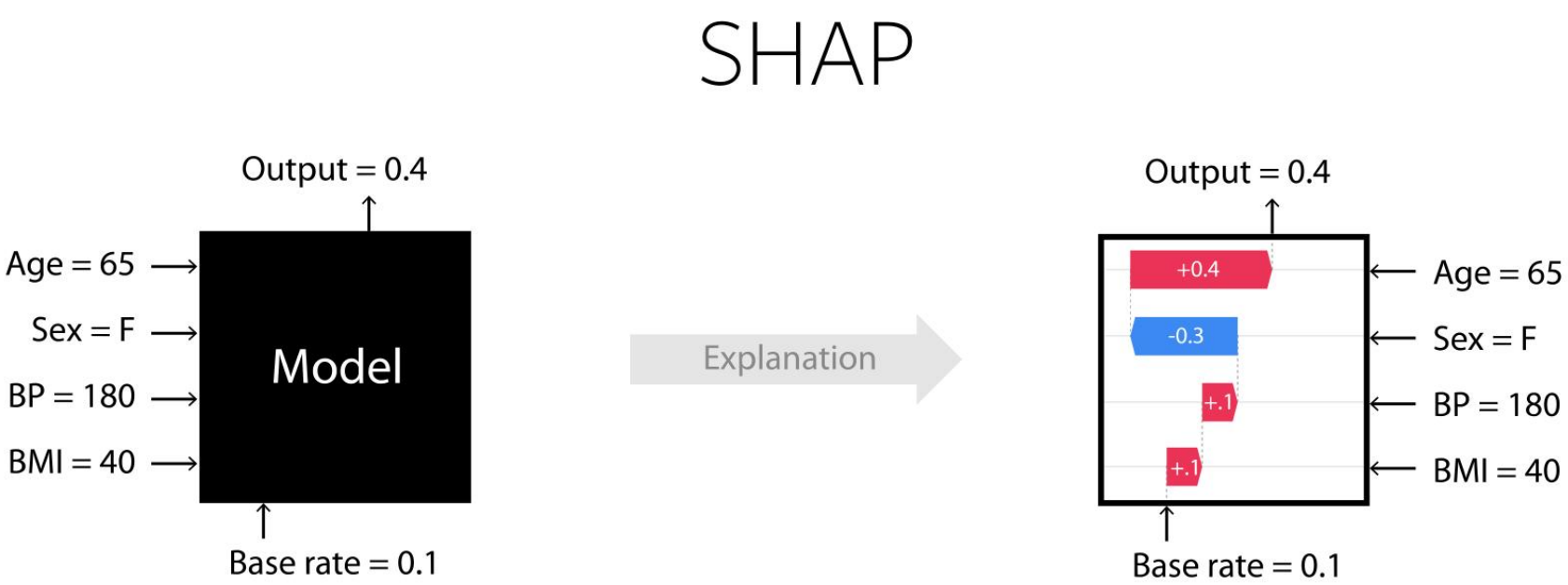
Feature importance



feature importance用来衡量数据集中每个特征的重要性。每个特征对于提升整个模型的预测能力的贡献程度就是特征的重要性。

- 这些因素与房价之间是正相关还是负相关?
- 这些因素之间是否存在交互影响?
- 每个特征对每个房价个体的影响如何?

SHAP简介



SHAP(SHapley Additive exPlanations):是一种博弈论方法，用于解释任何机器学习模型的输出。虽然来源于合作博弈论，但只是以该思想作为载体。在进行局部解释时，SHAP的核心是计算其中每个特征变量的Shapley Value。

Shapley：代表对每个样本中的每个特征变量，都计算出它的Shapley Value
Additive：代表对每一个样本而言，特征变量对应的Shapley Value是可加的
exPlanation：代表对单个样本的解释，即每个特征变量是如何影响模型的预测值

SHAP简介-Shapley Value

Shapley Value由加州大学教授罗伊德·夏普利（Lloyd Shapley）提出。夏普利值，来源于合作博弈理论，是一种基于贡献的分配方式。

今天加班，一个程序有500行代码需要编写，
产品经理找了三个程序猿来完成，按照完成量发奖金：
1号屌丝程序猿独立能写100行
2号大神程序猿独立能写125行
3号美女程序猿能写50行
1,2号合作能写270行
2,3号合作能写350行
1,3号合作能写375行
3个人共同能完成500行
但是，应该按照什么样的比例分配奖金呢？

概率	顺序	1号的边际贡献	2号的边际贡献	3号的边际贡献
1/6	1,2,3	100	170	230
1/6	1,3,2	100	125	275
1/6	2,1,3	145	125	230
1/6	2,3,1	150	125	225
1/6	3,1,2	325	125	50
1/6	3,2,1	150	300	50
Shapley Value		161.67	161.67	176.67
Ratio		32.33%	32.33%	35.33%

SHAP简介

SHAP构建一个加性的解释模型，所有的特征都视为“贡献者”。对于每个预测样本，模型都产生一个预测值，SHAP value就是该样本中每个特征所分配到的数值。

假设第i个样本为x_i，第i个样本的第j个特征为$x_{i,j}$，模型对第i个样本的预测值为y_i，整个模型的基线（基线值等于训练集的目标变量的拟合值的均值）为y_{base}，那么SHAP value服从以下等式。

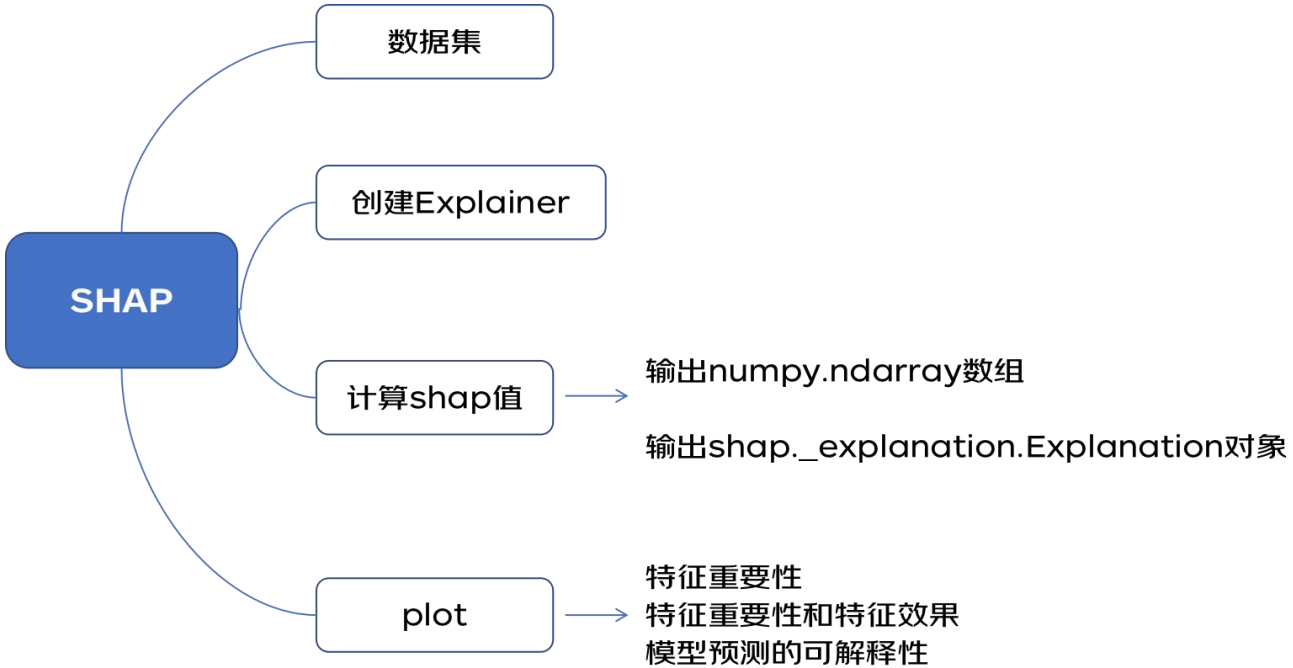
$yi=y_{base}+f(x_{i,1})+f(x_{i,2})+\cdots+f(x_{i,k})$，

其中$f(x_{i,j})$为$x_{i,j}$的SHAP值。

直观上看，$f(x_{i,1})$就是第i个样本中第1个特征对y_i的贡献值，当$f(x_{i,1})>0$，说明该特征提升了预测值，有正向作用；反之，说明该特征使得预测值降低，有反作用。

SHAP value最大的优势是SHAP能对于反映出每一个样本中的特征的影响力，而且还表现出影响的正负性。

SHAP简介

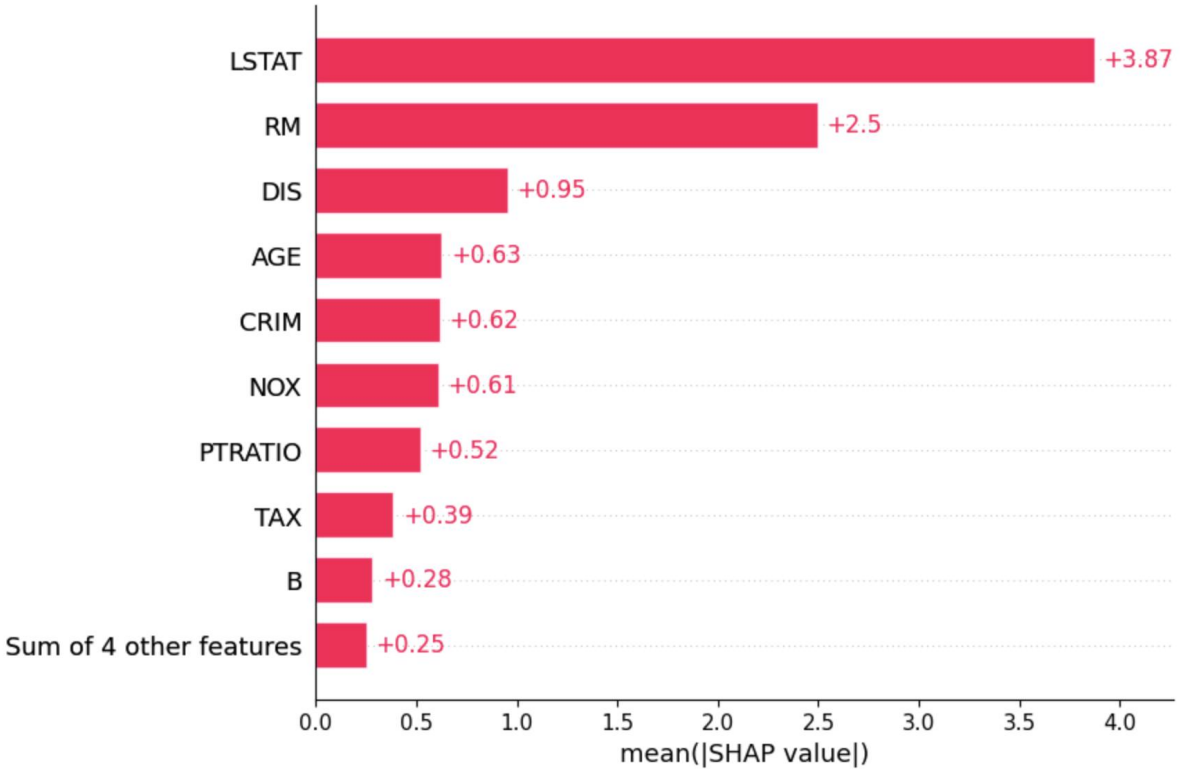



```python
import xgboost
import shap
import pandas as pd
import numpy as np

# 数据集
X, y = shap.datasets.boston()

# 创建Explainer
model = xgboost.XGBRegressor().fit(X, y)
explainer = shap.Explainer(model)

# 计算shap值
shap_values = explainer(X)

# plot
shap.plots.waterfall(shap_values[0])
```

SHAP的部分图形展示

总体特征图，针对整体特征重要性进行可视化展示

shap.plots.bar(shap_values)

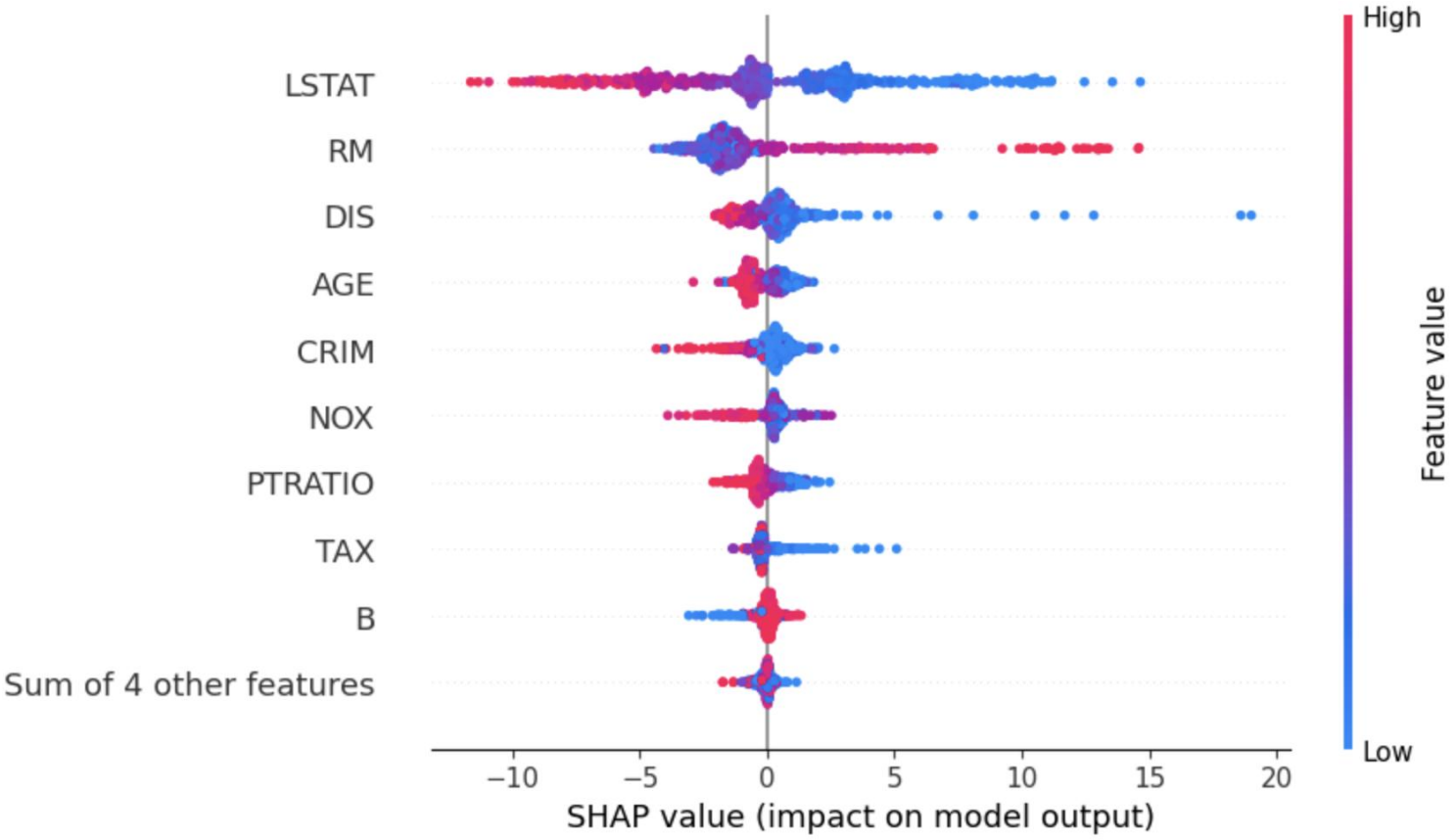


每个特征的全局重要性：每个样本中每个特征对应shap值绝对值取平均

SHAP的部分图形展示

总体特征图，针对整体特征重要性进行可视化展示

shap.summary_plot(shap_values, X)



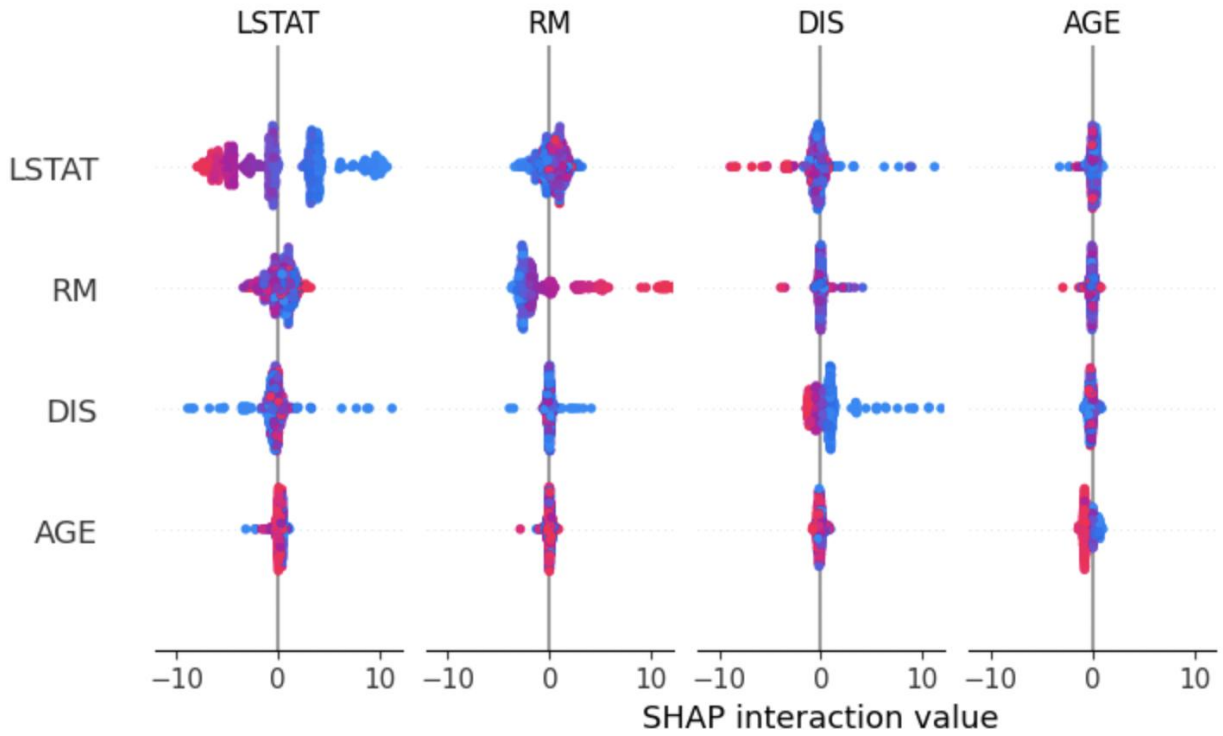
每一行代表一个特征，横坐标为SHAP值。

一个点代表一个样本，颜色表示特征值(红色高，蓝色低)。

比如，这张图表明LSTAT特征较高的取值会降低预测的房价

SHAP的部分图形展示

总体特征图，针对整体特征重要性进行可视化展示




```
# 对多个变量的交互进行分析
shap_interaction_values = explainer.shap_interaction_values(X)
shap.summary_plot(shap_interaction_values, X, max_display=4)
```

interaction value是将SHAP值推广到更高阶交互的一种方法。

```
print('第一个特征的shap values:', shap_values[0][0].values)
print('第一个特征的shap interaction values:', shap_interaction_values[0][0])
print('第一个特征的shap interaction values和:', shap_interaction_values[0][0].sum())
```

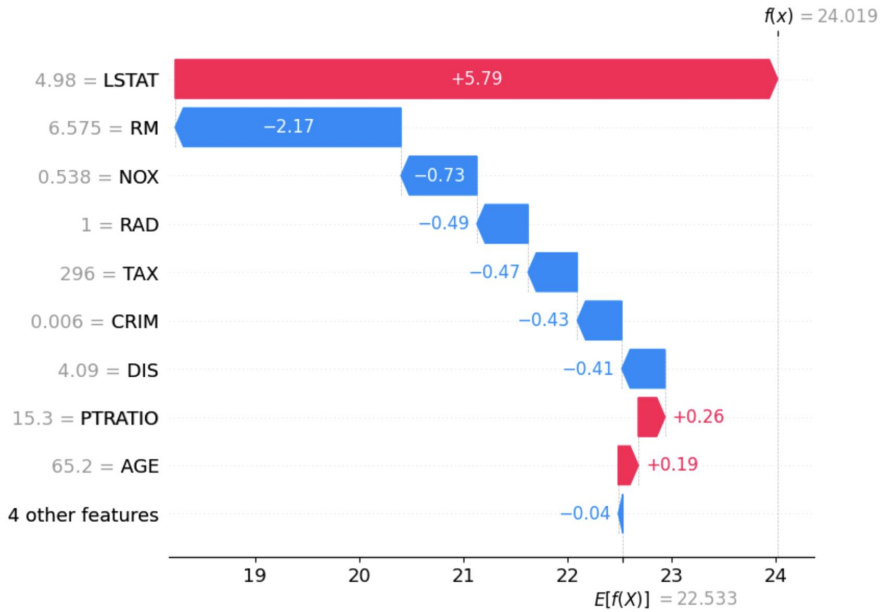
```
第一个特征的shap values: -0.42850167
第一个特征的shap interaction values: [ 0.2390606  -0.00276843 -0.00705554  0.00125731 -0.05565065  0.03624928
 -0.01305855 -0.05080049 -0.01404865  0.04045561 -0.15794721  0.00202929
 -0.4462242 ]
第一个特征的shap interaction values和: -0.42850167
```

SHAP的部分图形展示

针对单个样本进行可视化展示

shap.plots.force(shap_values[0])

shap.plots.waterfall(shap_values[0])

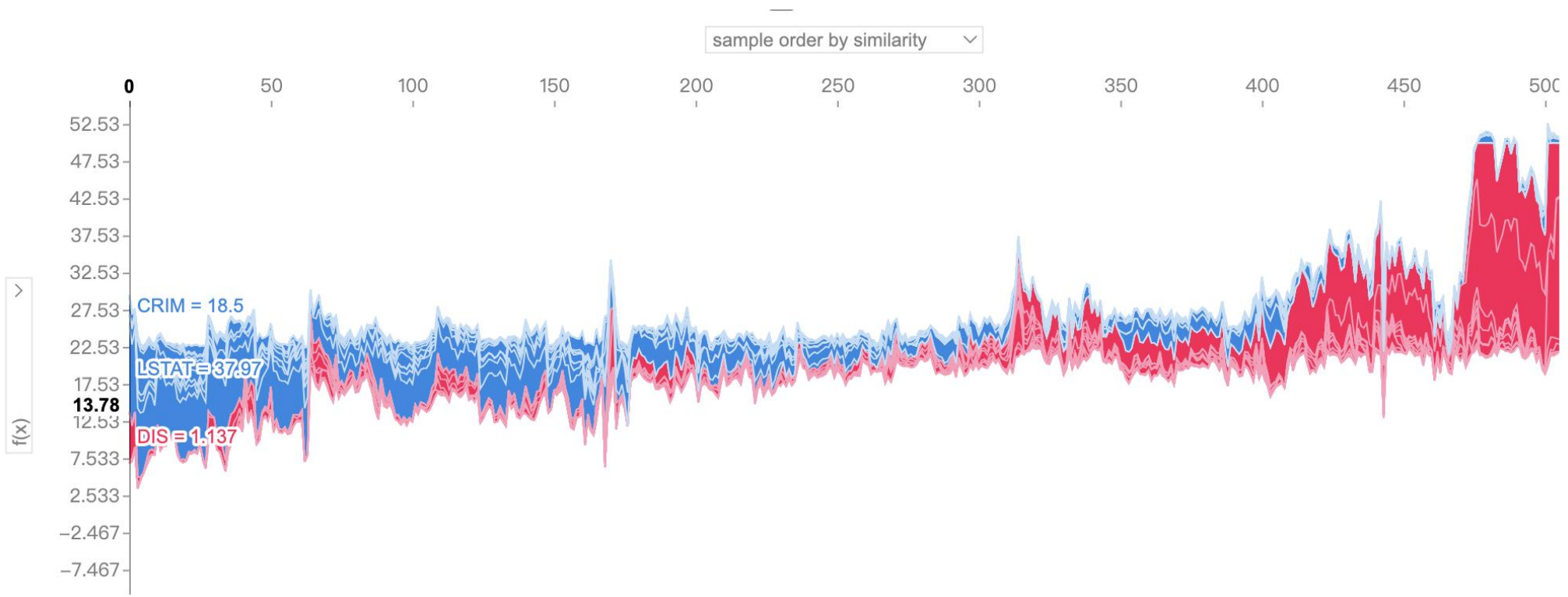


图中数据说明：

- f(x)为预测值
- E[f(x)]为基线值(base value)
- 灰色为样本特征值，图中为该样本该特征shap值
- 基线值等于训练集的目标变量的拟合值的均值
- 一个样本中各个特征SHAP的值的和加上基线值应该等于该样本的预测值

SHAP的部分图形展示

针对多个样本进行可视化展示 shap.plots.force(shap_values, X)

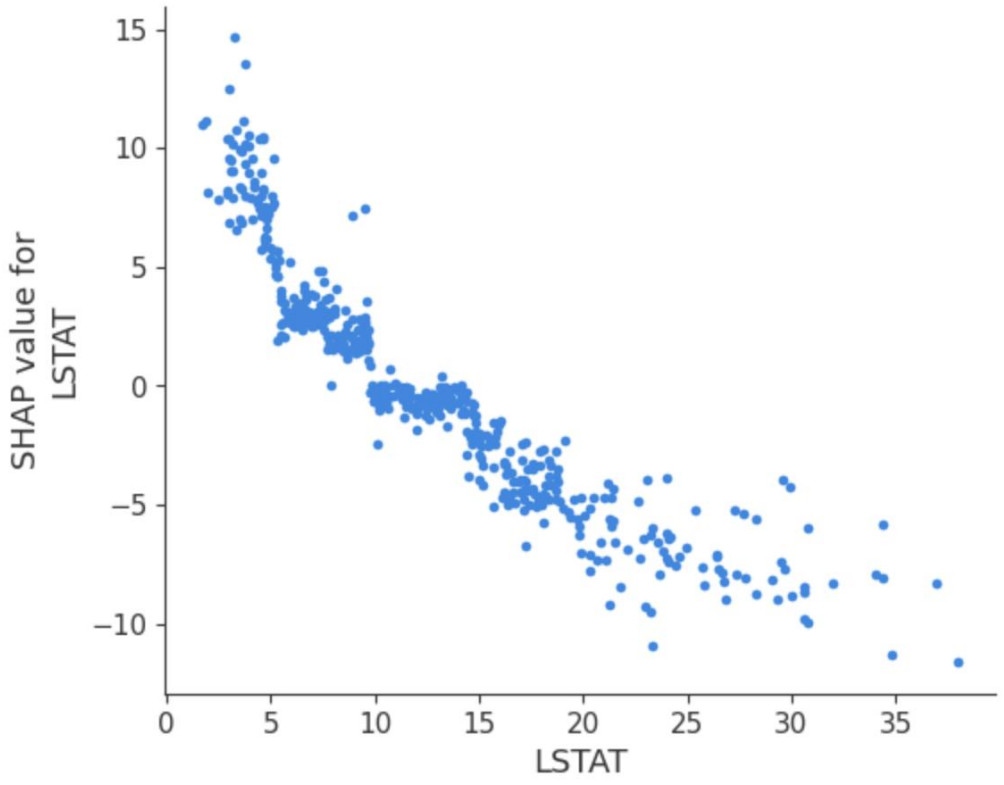


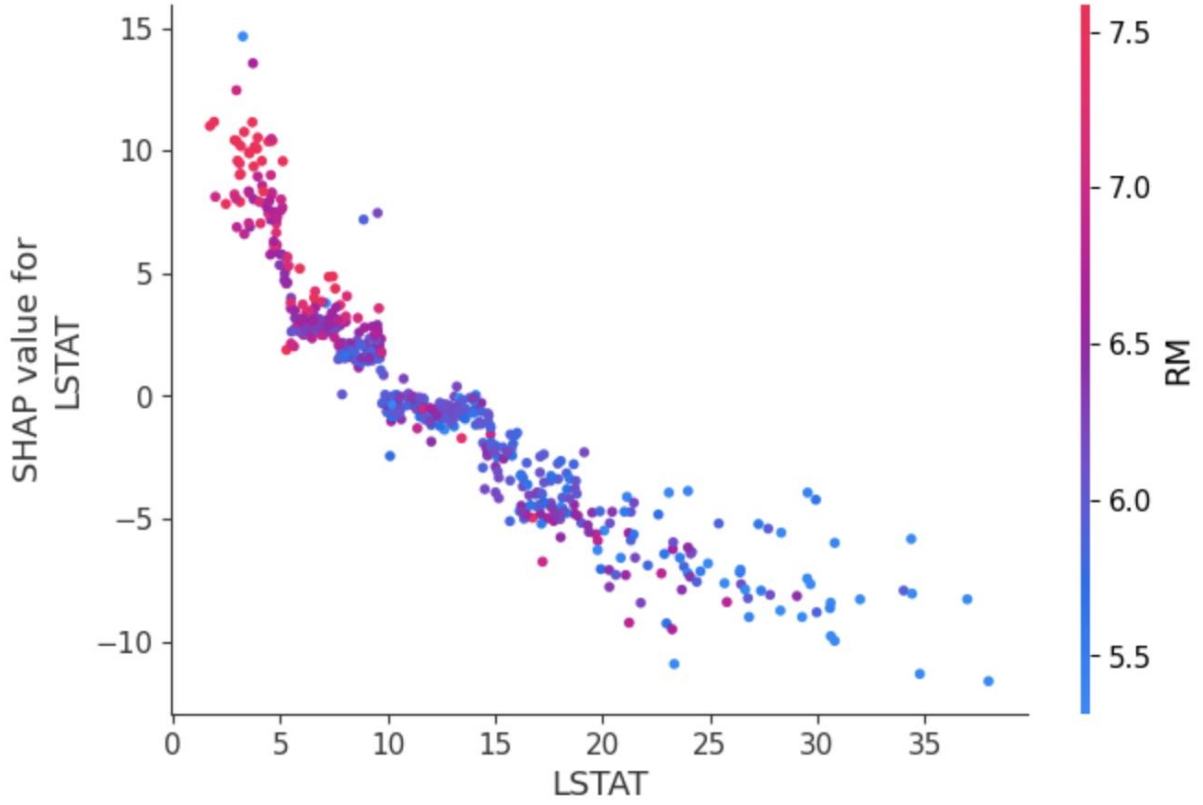
如果对多个样本进行解释，将单样本形式旋转90度然后水平并排放置，可以看到整个数据集的explanations

SHAP的部分图形展示

部分依赖图（Partial Dependence figure）

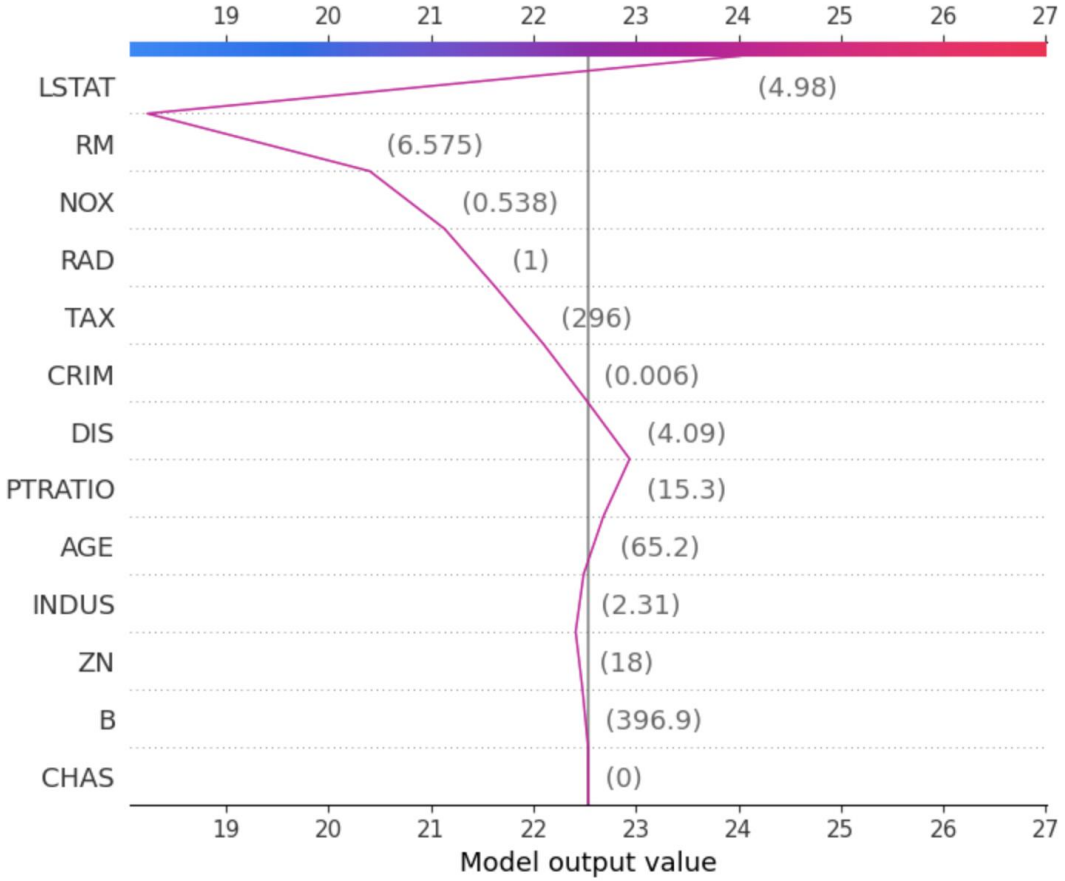
部分依赖图显示了一个或两个特征对机器学习模型的预测结果的边际效应。部分依赖图可以显示目标和特征之间的关系是线性的、单调的还是复杂的

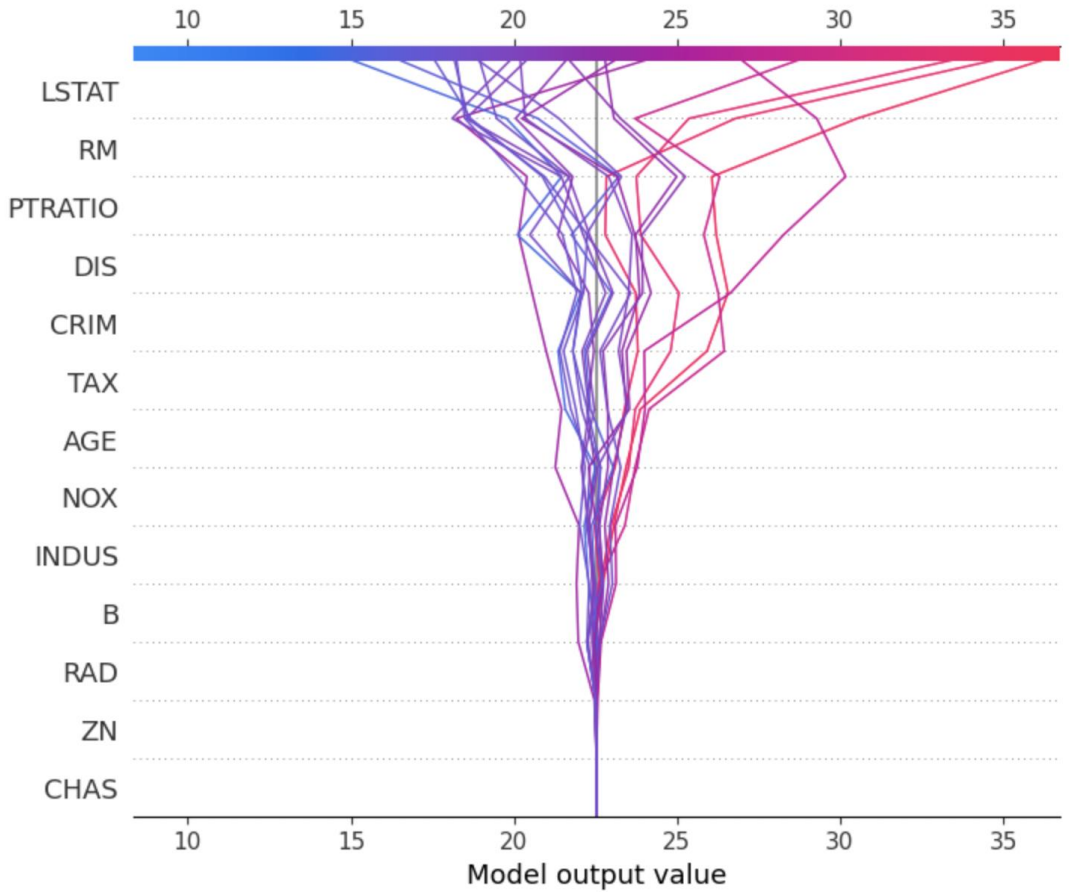




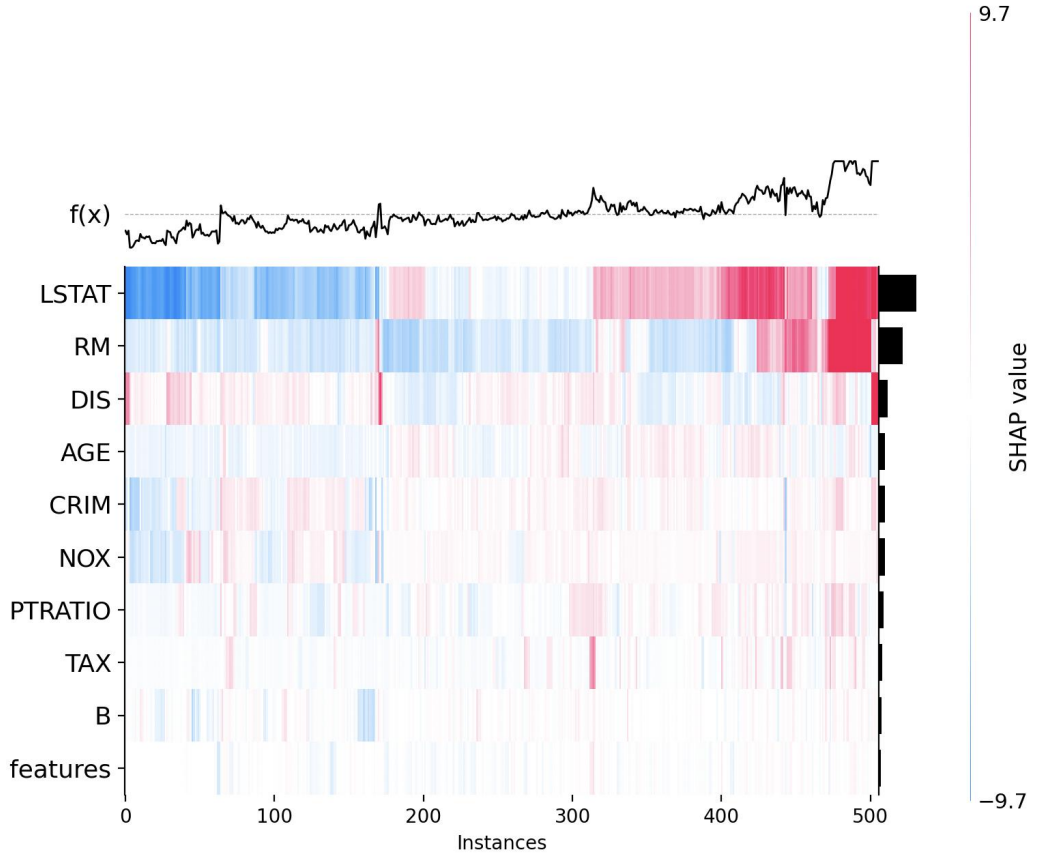
SHAP的部分图形展示

决策图中间灰色垂直直线标记了模型的基础值，彩色线是预测，表示每个特征是否将输出值移动到高于或低于平均预测的值。特征值在预测线旁边以供参考。从图的底部开始，预测线显示 SHAP value 如何从基础值累积到图顶部的模型最终分数





SHAP的部分图形展示



在热图矩阵上方是模型的输出，灰色虚线是基线(.base_value)，图右侧的条形图是每个模型输入的全局重要性

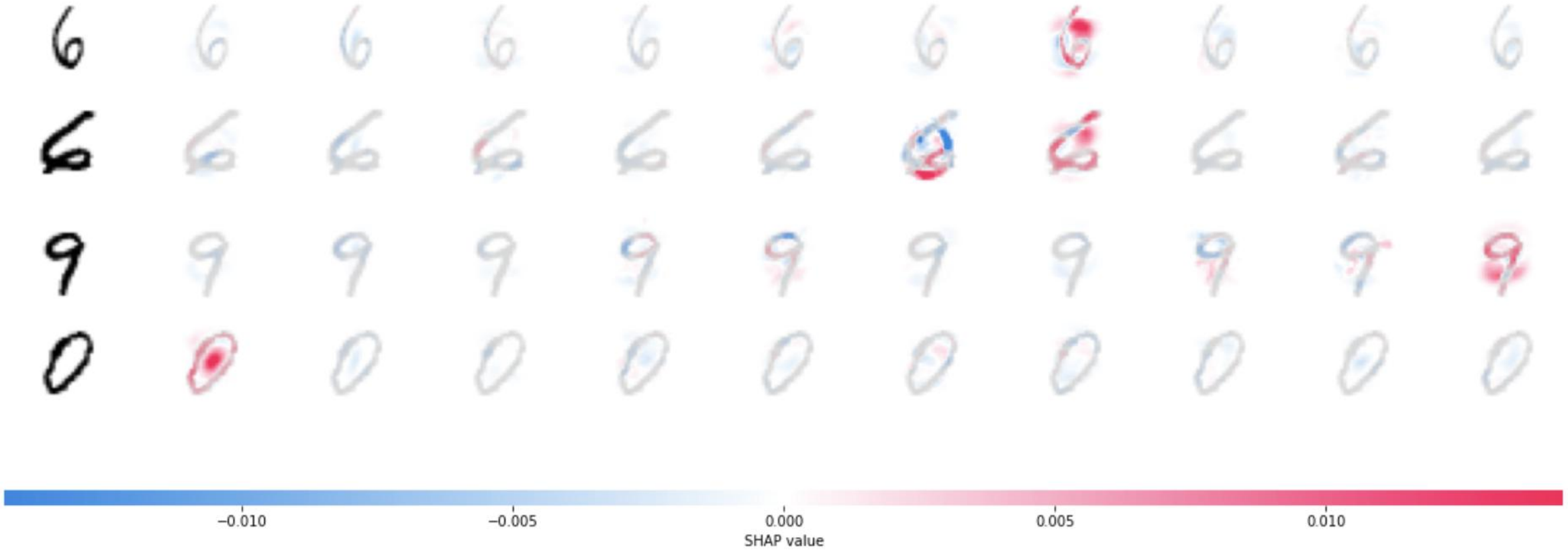
Text examples

Machine Translation Explanations
模型：Helsinki-NLP/opus-mt-en-zh

输入：Life was like a box of chocolates, you never know what you're gonna get

Image examples

PyTorch Deep Explainer MNIST example
MNIST 数据集：http://yann.lecun.com/exdb/mnist/



参考文档

Github：https://github.com/slundberg/shap/tree/master

shap：https://shap.readthedocs.io/en/latest/index.html

https://zhuanlan.zhihu.com/p/83412330

https://zhuanlan.zhihu.com/p/441302127

https://www.cnblogs.com/cgmcoding/p/15339638.html