*Дата последней сборки документа:*
*11 апреля 2020 г. в 20:42*

# Заметки по машинному обучению и анализу данных

*Подвойский А.О.*

Здесь приводятся заметки по некоторым вопросам, касающимся машинного обучения, анализа данных, программирования на языках `Python`, `R` и прочим сопряженным вопросам так или иначе, затрагивающим работу с данными.

## Содержание

# 1. Градиентный бустинг

## 1.1. Общие сведения

## 1.2. Особенности реализации в пакете sklearn

## 1.3. Особенности реализации в пакете XGBoost

### 1.3.1. Установка пакета xgboost на Windows

Устанавливать пакет **xgboost** рекомендуется с помощью следующей команды

```
conda install -c anaconda py-xgboost
```

Существует альтернативный способ установки пакета **xgboost** (разумеется он работает и для других пакетов). Для начала требуется вывести список доступных каналов (см. рис. 1), по которым будет проводиться поиск интересующего пакета (в данном случае пакета **xgboost**), а затем можно воспользоваться конструкцией

```
anaconda search -t conda xgboost
```

После, выбрав канал, можно приступать к установке пакета

```
conda install -c free py-xgboost
```

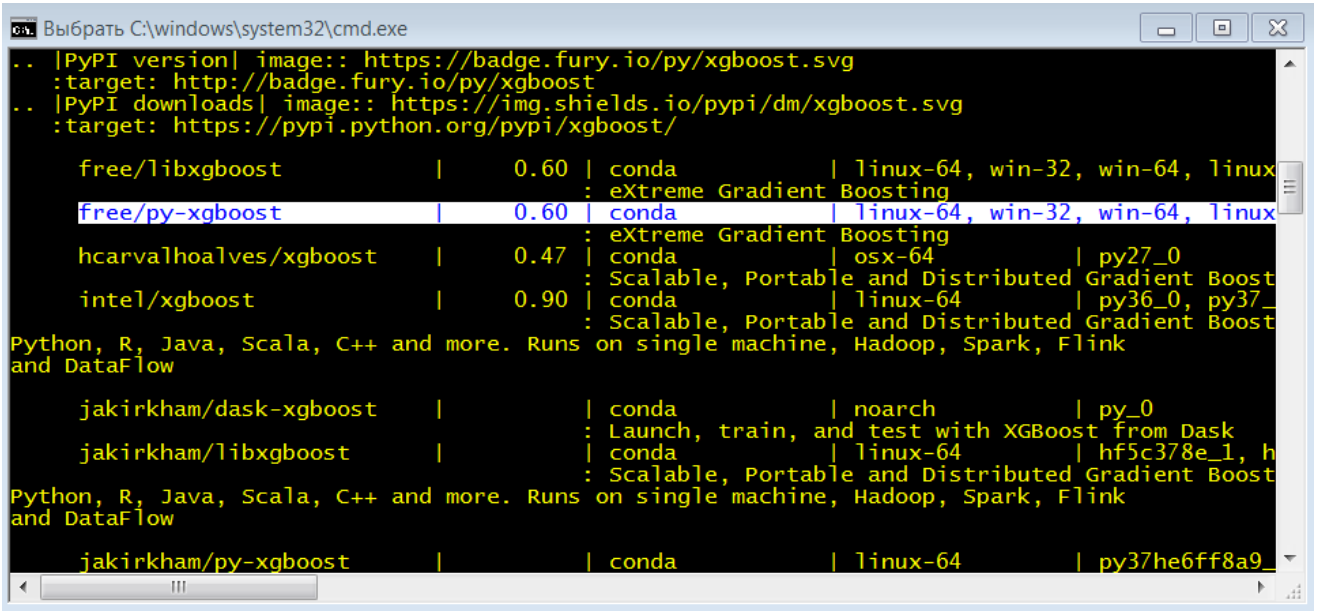



Рис. 1. Окно командной оболочки `cmd.exe` со списком доступных каналов, по которым будет проводиться поиск пакета `xgboost`

### 1.3.2. Простой пример работы с `xgboost` и `shap`

Решается задача бинарной классификации. Требуется построить модель, предсказывающую годовой доход заявителя по порогу \$50'000 (то есть больше или меньше \$50'000 зарабатывает заявитель в год). Используется набор данных UCI Adult income

```
import xgboost
import shap  # для оценки важности признаков вычисляются значения Шепли (Shapley value)
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split

shap.initjs()

X, y = shap.datasets.adult()
X_display, y_display = shap.datasets.adult(display=True)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=7)
d_train = xgboost.DMatrix(X_train, label=y_train)
d_test = xgboost.DMatrix(X_test, label=y_test)

params = {
    'eta' : 0.01,
    'objective' : 'binary:logistic',
    'subsample' : 0.5,
    'base_score' : np.mean(y_train),
    'eval_metric' : 'logloss'
}
model = xgboost.train(params, d_train,
                      num_boost_round = 5000,  # число итераций бустинга
                      evals = [(d_test, 'test')],
                      verbose_eval=100,  # выводит результат на каждой 100-ой итерации бустинга
                      early_stopping_rounds=20)

xgboost.plot_importance(model)
```

На рис. 2, рис. 3 и рис. 4 изображены графики важности признаков.

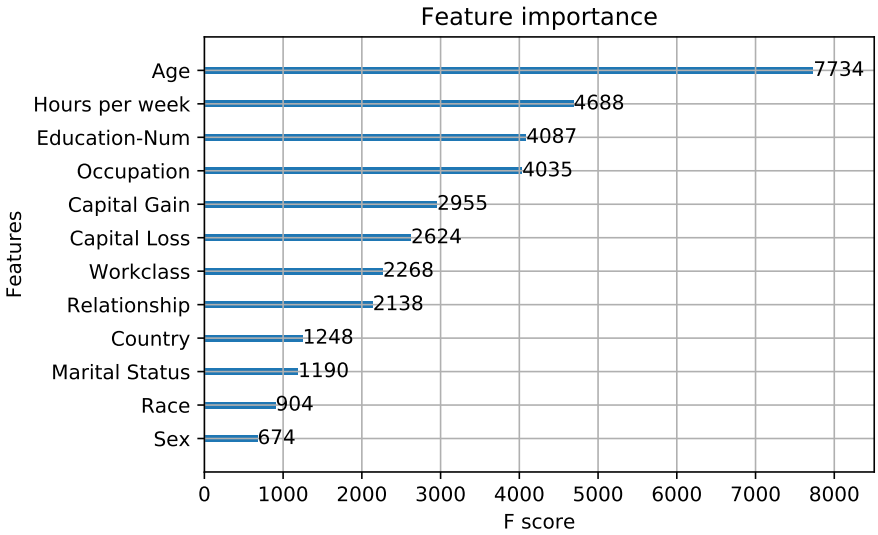



Рис. 2. График важности признаков `xgboost.plot_importance(model)`, построенный с помощью пакета `xgboost`

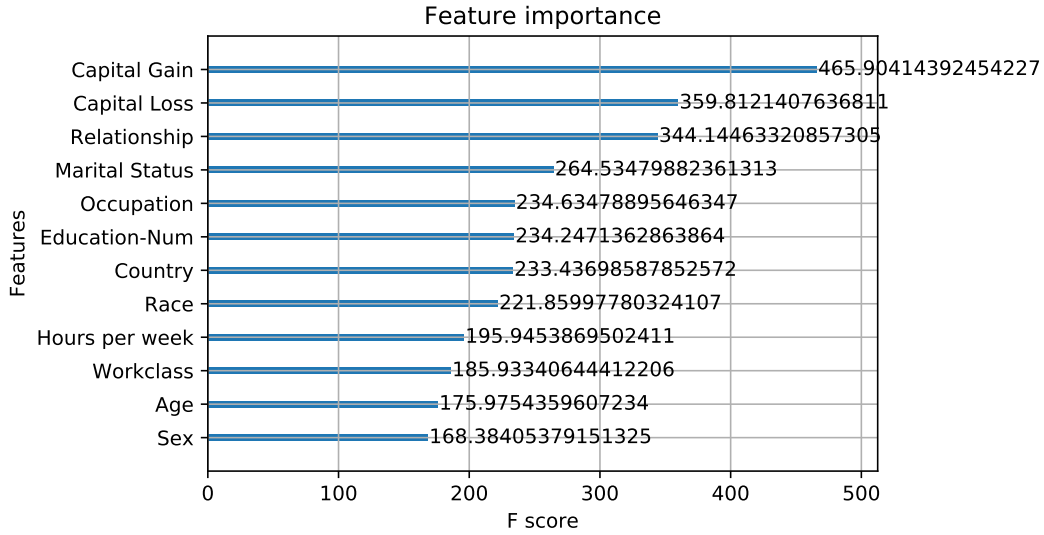



Рис. 3. График важности признаков `xgboost.plot_importance(model, importance_type='cover')`, построенный с помощью пакета `xgboost`

Следует иметь в виду, что в библиотеке `xgboost` поддерживается три варианта вычисления важности признаков (см. Interpretable Machine Learning with XGBoost):

- `weight`: общее число сценариев по всем деревьям, когда $i$-ый признак используется для расщепления обучающего набора данных,
- `cover`: общее число сценариев по всем деревьям, когда $i$-ый признак используется для расщепления набора данных, взвешенное по числу точек обучающего набора данных, которые проходят через эти расщепления,
- `gain`: среднее снижение потерь на обучающем наборе данных, полученное при использовании $i$-ого признака.

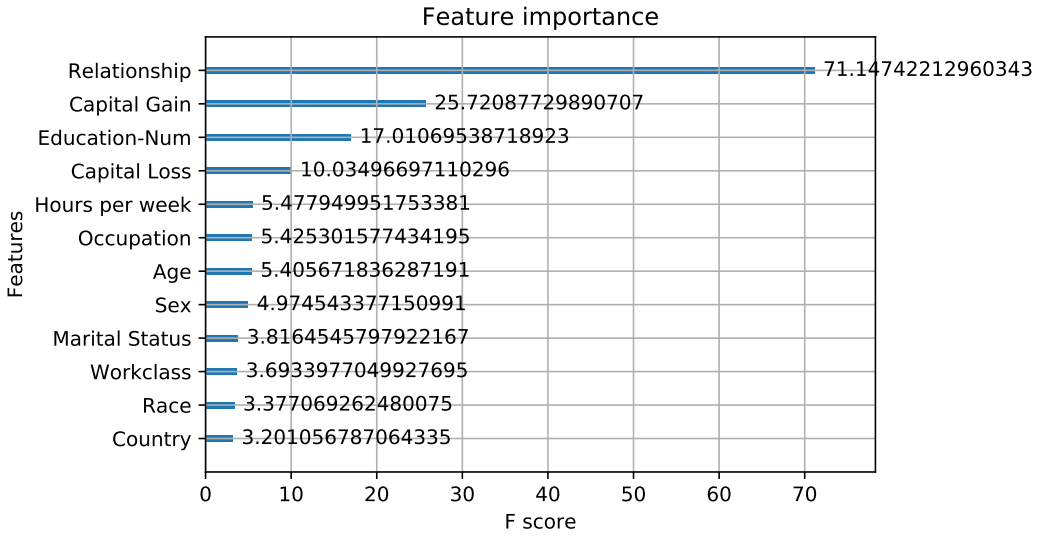



Рис. 4. График важности признаков `xgboost.plot_importance(model, importance_type='gain')`, построенный с помощью пакета `xgboost`

### 1.4. Особенности реализации в пакете `LightGBM`

### 1.5. Особенности реализации в пакете `CatBoost`

## 2. Приемы интерпретации моделей машинного обучения

## 3. Области видимости в языке `Python`

Когда мы говорим о поиске значения имени применительно к программному коду, под термином *область видимости* подразумевается *пространство имен* – то есть место в программном коде, где имени было присвоено значение [1].

В любом случае область видимости переменной (где она может использоваться) всегда определяется местом, где ей было присвоено значение.

---

*Замечание*

Термины «*область видимости*» и «*пространство имен*» можно использовать как синонимичные

---

При каждом вызове функции создается новое *локальное пространство имен*. Это пространство имен представляет локальное окружение, содержащее имена параметров функции, а также имена переменных, которым были присвоены значения в теле функции.

По умолчанию операция присваивания создает локальные имена (это поведение можно изменить с помощью `global` или `local`).

Схема разрешения имен в языке `Python` иногда называется *правилом LEGB*[1] [1, стр. 477]:

- Когда внутри функции выполняется обращение к неизвестному имени, интерпретатор пытается отыскать его в четырех областях видимости – в *локальной*, затем в *локальной области любой объемлющей функции* или в выражении `lambda`, затем в *глобальной* и, наконец, во *встроенной*. Поиск завершается, как только будет найдено первое подходящее имя.

[1] Local, Enclosing, Global, Built-in

- Когда внутри функции выполняется операция присваивания `a=10` (а не обращения к имени внутри выражения), интерпретатор всегда создает или изменяет имя в *локальной области видимости*, если в этой функции оно не было объявлено глобальным или нелокальным.

Пример

```
# глобальная область видимости
X = 99

def func(Y):  # Y и Z локальные переменные
    # локальная область видимости
    Z = X + Y # X - глобальная переменная
    return Z

func(1)  # Y = 1
```

Переменные `Y` и `Z` являются *локальными* (и существуют только во время выполнения функции), потому что присваивание значений обоим именам осуществляется внутри определения функции: присваивание переменной `Z` производится с помощью инструкции `=`, а `Y` – потому что аргументы всегда передаются через операцию присваивания.

Когда внутри функции выполняется операция присваивания значения переменной, она всегда выполняется в *локальном пространстве имен функции*

```
a = 10  # глобальная область видимости

def f():
    a = 100  # локальная область видимости
    return a
```

В результате переменная `a` в теле функции ссылается на совершенно другой объект, содержащий значение 100, а не тот, на который ссылается внешняя переменная.

Переменные во вложенных функциях привязаны к *лексической области видимости*. То есть поиск имени переменной начинается в *локальной области видимости* и затем последовательно продолжается во всех *объемлющих областях видимости внешних функций*, в направлении от внутренних к внешним.

Если и в этих *пространствах имен* искомое имя не будет найдено, поиск будет продолжен в *глобальном пространстве имен*, а затем во *встроенном пространстве имен*, как и прежде.

При обращении к локальной переменной до того, как ей будет присвоено значение, возбуждается исключение `UnboundLocalError`. Следующий пример демонстрирует один из возможных сценариев, когда такое исключение может возникнуть

```
i = 0
def foo():
    i = i + 1  # приведет к исключению UnboundLocalError
    print(i)
```

В этой функции переменная `i` определяется как *локальная* (потому что внутри функции ей присваивается некоторое значение и отсутствует инструкция `global`).

При этом инструкция присваивания `i = i + 1` пытается прочитать значение переменной `i` еще до того, как ей будет присвоено значение.

Хотя в этом примере существует глобальная переменная `i`, она не используется для получения значения. Переменные в функциях могут быть либо *локальными*, либо *глобальными* и не могут произвольно изменять *область видимости* в середине функции.

*Замечание*

Оператор `global` делает локальную переменную в теле функции *глобальной* и говорит интерпретатору чтобы тот не искал переменную в локальной области видимости текущей функции

Например, нельзя считать, что переменная `i` в выражении `i + 1` в предыдущем фрагменте обращается к глобальной переменной `i`; при этом переменная `i` в вызове `print(i)` подразумевает локальную переменную `i`, созданную в предыдущей инструкции.

*Обобщение по вопросу*

Когда интерпретатор, построчно сканируя тело функции `def`, натыкается на строку `i = i + 1`, он заключает что переменная `i` является *локальной*, так как ей присваивается значение именно в теле функции. А когда функция вызывается на выполнение и интерпретатор снова доходит до строки `i = i + 1`, выясняется, что переменная `i`, стоящая в правой части, не имет ссылок на какой-либо объект и потому возникает ошибка `UnboundLocalError`

# 4. Декораторы в Python

*Декораторы* выполняются *сразу после* загрузки или импорта модуля, однако увидеть какие-либо изменения можно только в том случае, если декоратор явно взаимодействует с пользователем на «верхнем уровне»[2], например, печатает строку в терминале. *Задекорированные* же функции выполняются строго в результате явного вызова [4, стр. 217].

## 4.1. Реализация простого декоратора

Рассмотрим простой декоратор, который хронометритует каждый вызов задекорированной функции и печатает затраченное время

clockdeco.py, не очень удачный пример декоратора

```
import time

def clock(func):
    print('test string from `clock`')  # <- строка будет выведена в терминал
                                       # сразу после загрузки модуля, который
                                       # импортирует данный декоратор
    def clocked(*args):  # замыкание
        t0 = time.perf_counter()  # запомнить начальный момент времени
        result = func(*args)  # вызвать функцию
        elapsed = time.perf_counter() - t0  # вычислить сколько прошло времени
        name = func.__name__
        arg_str = ', '.join(repr(arg) for arg in args)
        print(f'{elapsed}, {name}({arg_str}) -> {result}')
        return result  # вернуть результат
return clocked
```

Использование декоратора выглядит так

clockdeco_demo.py

```
1 import time
2 from clockdeco import clock
```

[2]Если декоратор простой одноуровневый, то под верхним уровнем понимается его локальная область видимости, а если декоратор содержит замыкание, то – понимается область видимости объемлющей функции

```python

def simple_deco_1(f):
    '''
    Декоратор с замыканием
    '''
    def inner():
        print('test string from `simple_deco_1`')  # <- строка НЕ будет выведена
                                                   # после загрузке модуля
    return inner

def simple_deco_2(f):
    '''
    Простой одноуровневый декоратор
    '''
    print('test string from `simple_deco_2`')  # <- строка будет выведена в терминал
                                               # сразу после загрузки модуля
    return f

@simple_deco_1  # simple_func_1 = simple_deco_1(f=simple_func_1) -> inner
def simple_func_1():
    print('test string from `simple_func_1`')

@simple_deco_2  # simple_func_2 = simple_deco_2(f=simple_func_2) -> simple_func_2
def simple_func_2():
    print('test string from `simple_func_2`')

@clock  # snooze = clock(func=snooze) -> clocked
def snooze(seconds):
    time.sleep(seconds)

@clock
def factorial(n):
    return 1 if n < 2 else n*factorial(n-1)


if __name__ == '__main__':
    print('*'*10, 'Calling snooze(.123)')
    print('snooze_result = {}'.format(snooze(.123)))
    print('*'*10, 'Calling factorial(6)')
    print('6! = ', factorial(6))
    print(f'This is result from `simple_func_1`: {simple_func_1()}')
    print(f'This is result from `simple_func_2`: {simple_func_2()}')
```

Вывод clockdeco_demo.py

```
test string from `simple_deco_2`
test string from `clock`
test string from `clock`
********** Calling snooze(.123)
0.1261, snooze(0.123) -> None
snooze_result = None
********** Calling factorial(6)
1.866e-06, factorial(1) -> 1
7.589e-05, factorial(2) -> 2
0.0001266, factorial(3) -> 6
0.0001732, factorial(4) -> 24
0.0002224, factorial(5) -> 120
0.0002715, factorial(6) -> 720
6! =  720
test string from `simple_deco_1`
```

```
this is result from ‘simple_func_1‘: None
test string from ‘simple_func_2‘
this is result from ‘simple_func_2‘: None
```

---

*Замечание*

Приведенный выше пример декоратора `clock` из модуля `clockdeco.py` не удачен в том смысле, что если нам, например, потребуется вывести значение атрибута `__name__` задекорированной функции `snooze`, т.е. `snooze.__name__`, то будет возвращена строка `'clocked'`, а не `'snooze'`.

Чтобы декоратор «не портил» значения атрибута `__name__`, следует задекорировать замыкание декоратора с помощью `@functools.wraps(func)`

---

При разгрузке модуля `clockdeco_demo.py` будут выполнены все декораторы, но только декораторы `simple_deco_2` и `clock` выведут в терминал строки, потому как эти строки расположены на верхнем уровне декораторов (т.е. находятся не внутри вложенных функций). Декоратор `simple_deco_1` ничего не выводит, так как строка находится в области видимости вложенной функции.

Важно отметить следующее: после загрузки модуля, как уже говорилось выше, будут выведены в терминал строки, расположенные на верхнем уровне декораторов, но самое главное заключается в том, что после выполнения декоратора `clock` объект `snooze` уже будет ссылаться на внутреннюю функцию `clocked` декоратора `clock`, а после выполнения декоратора `simple_deco_1` объект `simple_func_1` будет ссылаться на внутреннюю функцию `inner`. Что же касается декоратора `simple_deco_2`, то объект `simple_func_2` будет ссылаться на `simple_func_2`.

По этой причине при вызове функции `simple_func_1()` печатается строка из внутренней функции `inner`, а при вызове функции `simple_func_2()` – строка из этой же функции.

Еще один пример декоратора с замыканием

```
def deco(f):
    def inner(*args, **kwargs):
        print(f'from ‘deco-inner‘: args={args}, kwargs={kwargs}')
        return f  # f - свободная переменная
    return inner

@deco  # target = deco(f=target) -> inner :: target -> inner :: target=inner
def target(a, b=10):
    return (f'from ‘target‘: a={a}, b={b}')

print(target(20, b=500)(250))  # сначала вызывается inner(20, b=500), а потом target(250)
```

Выведет

```
from ‘deco-inner‘: args=(20,), kwargs={'b': 500}
from ‘target‘: a=250, b=10
```

## 4.2. Кэширование с помощью `functools.lru_cache`

Декоратор `functools.lru_cache` очень полезен на практике. Он реализует запоминание: прием оптимизации, смысл которого заключается в сохранении результатов предыдущих дорогостоящих вызовов функции, что позволяет избежать повторного вычисления с теми же аргументами, что и раньше [4, стр. 230].

Например

```
import functools
from clockdeco import clock

@functools.lru_cache
@clock
def fibonacci(n):
    if n < 2:
        return n
    return fibonacci(n-2) + fibonacci(n-1)

if __name__ == '__main__':
    print(fibonacci(6))
```

*Замечание*

`lru_cache` хранит результаты в словаре, ключи которого составлены из позиционных и именованных аргументов вызовов, а это значит, что все аргументы, принимаемые декорируемой функции должны быть *хешируемыми*

### 4.3. Одиночная диспетчеризация и обобщенные функции

Декоратор `functools.singledispatch` позволяет каждому модулю вносить свой вклад в общее решение. Обычная функция, декорированная `@singledispatch` становится *обобщенной функцией*: групповой функцией, выполняющей одну и ту же логическую операцию по-разному в зависимости от типа первого аргумента [4, стр. 234]. Именно это и называется *одиночной диспетчеризацией*. Если бы для выбора конкретных функций использовалось больше аргументов, то мы имели бы дело с *множественной диспетчеризацией*.

Например

```
from functools import singledispatch
from collections import abc
import numbers
import html

@singledispatch  # делает функцию обобщенной
def htmlize(obj):
    content = html.escape(repr(obj))
    return '<pre>{}</pre>'.format(content)


@htmlize.register(str)  # будет вызываться для объектов строкового типа данных
def _(text):
    content = html.escape(text).replace('\n', '<br>\n')
    return '<p>{}</p>'.format(content)


@htmlize.register(numbers.Integral)  # будет вызваться для объектов целочисленного типа данных
def _(n):
    return '<pre>{} (0x{:x})</pre>'.format(n)


@htmlize.register(tuple)
@htmlize.register(abc.MutableSequence)
def _(seq):
    inner = '</li>\n<li>'.join(htmlize(item) for item in seq)
```

```
    return '<ul>\n<li>' + inner + '</li>\n</ul>'
```

*Замечание*

По возможности следует стараться регистрировать специализированные функции для обработки абстрактных базовых классов, например, `numbers.Integral` или `abc.MutableSequence`, а не конкретные реализации типа `int` или `list`

Замечательное свойство механизма `singledispatch` состоит в том, что специализированные функции можно зарегистрировать в любом месте системы, в любом модуле [4].

### 4.4. Композиции декораторов

Когда два декоратора `@d1` и `@d2` применяются к одной и той же функции `f` в указанном порядке, получается то же самое, что в результате композиции `f = d1(d2(f))`.

Иными словами

```
@d1
@d2
def f():
    print('f')
```

эквивалентен следующему

```
def f():
    print('f')

f = d1(d2(f))
```

Рассмотрим еще один пример композиции декораторов

```
def deco1(f):  # выполняется вторым
    print('deco-1')  # # будет выведена в терминал
    def inner1():
        print('string from `deco1-inner``')
    return inner1

def deco2(f):  # выполняется первым
    print('deco-2')  # будет выведена в терминал
    def inner2():
        print('string from `deco2-inner`')
    return inner2

@deco1  # 2) inner2 = deco1(f=inner2) -> inner1 :: inner2 -> inner1 :: inner2 = inner1
@deco2  # 1) target = deco2(f=target) -> inner2 :: target -> inner2 :: target = inner2
def target():  # 3) target -> inner1
    print('string from `target``')


if __name__ == '__main__':
    target()  # выведет string from `deco1-inner`
```

Выведет

```
deco-2
deco-1
string from `deco1-inner`
```

*Замечание*

Первым выполняется тот декоратор, который ближе расположен к декорируемой функции

То есть при загрузке или импорте модуля будут выполнены декораторы `deco1` и `deco2`: сначала `deco2`, а затем `deco1`, потому как `deco2` ближе к декорируемой функции. Декоратор `deco1` применяется к той функции, которую возвращает `deco2`.

## 4.5. Параметризованные декораторы

Параметризованные декораторы часто называют *фабриками декораторов*. Фабрики декораторов возвращают настоящие декораторы, которые применяются к декорируемой функции.

Пример

```
registry = set()

def register(activate=True):  # фабрика декораторов
    def decorate(func):  # декоратор
        print(f'running register(activate={activate})->decorate({func})')
        if activate:
            registry.add(func)
        else:
            registry.discard(func)
        return func
    return decorate

@register(activate=False)  # f1 = decorate(func=f1) -> f1 :: f1 -> f1
def f1():
    print('running f1()')

@register()  # f2 = decorate(func=f2) -> f2 :: f2 -> f2
def f2():
    print('running f2()')

def f3():
print('running f3()')
```

Идея в том, что функция `register()` возвращает декоратор `decorate`, который затем применяется к декорируемой функции [4].

*Замечание*

*Фабрика декораторов* возвращает *декоратор*, который применяется к декорируемой функции

Чуть подробнее: сразу после загрузки или импорта модуля выполняется фабрика декораторов `register`, которая возвращает декоратор `decorate`, который и применяется к функциям. Можно представлять, что фабрика декораторов нужна только для того, чтобы собрать значения каких-то дополнительных переменных, которые потребуются позже. В данном примере можно представить, что строка `@register()` заменяется на строку `@decorate`. То есть декоратор применяется к функции, расположенной на следующей строке, и работает как обычно.

Как можно работать с этой фабрикой декораторов

```
register()(f3)  # добавить ссылку на функцию f3 во множество registry
register(activate=False)(f2)  # удалить ссылку на функцию f2
```

Конструкция `register()` возвращает декоратор, который затем применяется к переменной (например, к `f3`), ассоциированной с декорируемой функцией, и работает так, как если бы изначально был только он (без фабрики декораторов) [4].

Если бы у декоратора был еще один уровень вложенности, т.е. было бы определено еще и замыкание, то это изменило бы только ссылку на функцию, которую возвращает замыкание

```
def fabricdeco():  # фабрика декораторов
    def deco(f):  # декоратор
        def inner():  # замыкание
            print(f'from inner: {f}')
        return inner
    return deco

@fabicdeco()  # target = deco(f=target) -> inner :: target -> inner :: target=inner
def target():
    print('from target')

target()  # на самом деле вызывается inner() -> from inner: <function target at 0x0...08B05318>
```

Рассмотрим еще один пример параметризованного декоратора

```
import time

DEFAULT_FMT = '[{elapsed}s] {name}({args}) -> {result}'

def clock(fmt=DEFAULT_FMT):  # фабрика декораторов
    def decorate(func):  # декоратор
        count = 0
        def clocked(*_args):  # замыкание
            nonlocal count  # делает переменную свободной
            count += 1
            print(f'args-{count}: {_args}')
            t0 = time.time()
            _result = func(*_args)
            elapsed = time.time() - t0
            name = func.__name__
            args = ', '.join(repr(arg) for arg in _args)
            result = repr(_result)
            print(fmt.format(**locals())) # использование **locals() позволяет ссылаться
                                          # на любую локальную переменную clocked
            return _result
        return clocked
    return decorate


if __name__ == '__main__':
    @clock()  # snooze = decorate(func=snooze) -> clocked :: snooze -> clocked
    def snooze(seconds):
        time.sleep(seconds)

    for i in range(3):
        snooze(0.123)
```

Теперь фабрику декораторов можно вызывать, например, так:

```
@clock('log::{name}({args}), dt={elapsed:.5g}s')
def snooze(seconds):
    time.sleep(seconds)
```

Объяснение: сразу после загрузки модуля (когда модуль загружается как скрипт), интерпретатор наталкивается на строку `@clock()` после чего вызывает *фабрику декораторов* `clock`, которая возвращает ссылку на *декоратор* `decorate`, который в свою очередь начинает работать как и в описанных выше случаях, т.е. аргумент `func` декоратора получает ссылку на `snooze`, а сам декоратор возвращает ссылку на *замыкание* `clocked`.

---

*Замечание*

Интерпретатор вызывает декоратор или фабрику декораторов из той строки, в которой находится конструкция `@deco`, поэтому как в данном примере если `@clock()` разместить в блоке проверки значения атрибута `__name__`, то фабрика декораторов не будет вызвана

---

Однако здесь есть любопытный момент. Переменные `func` и `count` вообще говоря являются *свободными переменными*, поэтому их значения можно читать из-под замыкания (находясь в области видимости замыкания) даже после того, как локальная область видимости объемлющей функции (декоратора) будет уничтожена. Но если попытаться передать новое значение переменной `count`, находясь в теле замыкания *без* использования оператора `nonlocal`, то это приведет к ошибке `UnboundLocalError`. Дело в том что свободные переменные по умолчанию можно только читать из-под замыкания. Когда мы присваиваем новое значение переменной `count` в теле замыкания, то мы делаем эту переменную локальной для замыкания `clocked`, переменная `count` перестает быть свободной. Чтобы объяснить интерпретатору, что переменная `count` должна рассматриваться как свободная даже если ей присваивается значение в области видимости замыкания, следует использовать оператор `nonlocal`.

## 4.6. Обобщение по механизму работы декораторов

Если обобщить сказанное выше, то получается, что задекорированная функция ссылается на ту функцию, которую возвращает декоратор, аргумент которого получил ссылку на данную функцию. И происходит это *сразу после* загрузки или импорта модуля. А затем остается только вызвать задекорированную функцию, которая вообще говоря уже ссылается на какую-то другую функцию, которую возвращает декоратор, т.е. если

```
def deco(f):
    def inner():  # замыкание
        print('inner')
    return inner

@deco  # выполняется при загрузке/импорте модуля
def target():
    print('target')
```

то `target = deco(f=target) -> inner`

и, следовательно, `target -> inner` (можно считать, что `target=inner`);

поэтому при вызове `target()` на самом деле вызывается `inner()` и будет выведена строка `'inner'` (см. рис. 5).

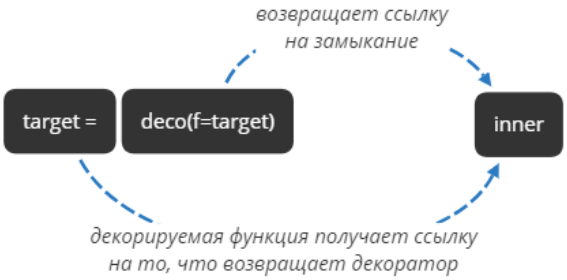



Рис. 5. К вопросу о механизме работы декоратора с вложенной функцией

## 5. Замыкания/фабричные функции в Python

Под термином *замыкание* или *фабричная функция* подразумевается объект функции, который сохраняет значения в *объемлющих областях видимости*, даже когда эти области могут прекратить свое существование [1, стр. 488].

В источнике [4, стр. 222] приводится несколько отличное определение[3]: *замыкание* – это вложенная функция с расширенной областью видимости, которая охватывает все *неглобальные* переменные, объявленные в области видимости объемлющей функции, и способная работать с этими переменными даже после того как локальная область видимости объемлющей функции будет уничтожена.

Замыкания и вложенные функции особенно удобны, когда требуется реализовать концепцию отложенных вычислений [2].

---

*Замечание*

Все же правильнее «фабрикой функций» называть всю конструкцию из объемлющей и вложенной функций, а «замыканием» – только вложенную функцию

---

Рассмотрим в качестве примера следующую функцию

```
def maker(N):
    def action(X):
        return X**N  # функция action запоминает значение N в объемлющей области видимости
    return action
```

Здесь определяется внешняя функция, которая просто создает и возвращает вложенную функцию, не вызывая ее. Если вызвать внешнюю функцию

```
>>> f = maker(2)  # запишет 2 в N
>>> f  # <function action at 0x0147280>
```

она вернет ссылку на созданную ею вложенную функцию, созданную при выполнении вложенной инструкции `def`. Если теперь вызвать то, что было получено от внешней функции

```
>>> f(3)  # запишет 3 в X, в N по-прежнему хранится число 2
>>> f(4)  # 4**2
```

будет вызвана вложенная функция, с именем `action` внутри функции `maker`. Самое необычное здесь то, что вложенная функция продолжает хранить число 2, значение переменной `N` в функции `maker` даже при том, что к моменту вызова функции `action` функция `maker` уже *завершила свою работу и вернула управление.*

[3] Определение содержит авторские правки

Когда функция используется как вложенная, в замыкание включается все ее окружение, необходимое для работы внутренней функции [2, стр. 137].

## 5.1. Области видимости и значения по умолчанию применительно к переменным цикла

Существует одна известная особенность для функций или lambda-выражений: если lambda-выражение или инструкция `def` вложены в цикл внутри другой функции и вложенная функция ссылается на переменную из объемлющей области видимости, которая изменяется в цикле, все функции, созданные в этом цикле, будут иметь одно и то же значение – значение, которое имела переменная на последней итерации [1, стр. 492].

Например, ниже предпринята попытка создать список функций, каждая из которых запоминает текущее значение переменной `i` из объемлющей области видимости

Эта реализация работать НЕ будет

```
def makeActions():
    acts = []
    for i in range(5):  # область видимости объемлющей функции
        acts.append(
            lambda x: i**x  # локальная область видимости вложенной анонимной функции
        )
    return acts

acts = makeActions()
print(acts[0](2))  # вернет 4**2, последнее значение i
print(acts[3](2))  # вернет 4**2, последнее значение i
```

Такой подход не дает желаемого результата, потому что поиск переменной в объемлющей области видимости производится позднее, *при вызове вложенных функций*, в результате все они получат одно и то же значение (значение, которое имела переменная цикла на последней итерации).

Это один из случаев, когда необходимо явно сохранять значение из объемлющей области видимости в виде аргумента со значением по умолчанию вместо использования ссылки на переменную из объемлющей области видимости.

То есть, чтобы фрагмент заработал, необходимо передать текущее значение переменной из объемлющей области видимости в виде значения по умолчанию. Значения по умолчанию вычисляются в момент *создания вложенной функции* (а не когда она *вызывается*), поэтому каждая из них сохранит свое собственное значение `i`

Правильная реализация

```
def makeActions():
    acts = []
    for i in range(5):
        acts.append(
            lambda x, i=i: i**x  # сохранить текущее значение i
        )
    return acts

acts = makeActions()
print(acts[0](2))  # вернет 0**2
print(acts[2](2))  # вернет 2**2
```

*Обобщение по вопросу*

Значения аргументов по умолчанию вложенных функций, динамически создаваемых в цикле на уровне области видимости объемлющей функции, вычисляются в момент *создания* этих вложенных функций, а не в момент их вызова, поэтому `lambda x, i=i: ...` работает корректно

# 6. Калибровка классификаторов

Подробности в статье А. Дьяконова «Проблема калибровки уверенности».

Ниже описываются способы оценить качество калибровки алгоритма. Надо сравнить *уверенность* (confidence) и *долю верных ответов* (accuracy) на тестовой выборке.

Если классификатор «хорошо откалиброван» и для большой группы объектов этот классификатор возвращает вероятность принадлежности к положительному классу 0.8, то среди этих объектов будет приблизительно 80% объектов, которые в действительности принадлежат положительному классу. То есть, если для группы точек данных общим числом 100 классификатор возвращает вероятность положительного класса 0.8, то приблизительно 80 точек на самом деле будут принадлежать положительному классу и доля верных ответов тогда составит 0.8.

## 6.1. Непараметрический метод гистограммной калибровки (Histogram Binning)

Изначально в методе использовались бины одинаковой ширины, но можно использовать и равномощные бины.

Недостатки подхода:

- число бинов задается наперед,
- функция деформации не непрерывна,
- в «равноширинном варианте» в некоторых бинах может содержаться недостаточное число точек.

Метод был предложен Zadrozny B. и Elkan C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers.

## 6.2. Непараметрический метод изотонической регрессии (Isotonic Regression)

Строится монотонно неубывающая функция деформации оценок алгоритма.

Метод был предложен Zadrozny B. и Elkan C. Transforming classifier scores into accurate multiclass probability estimates.

Функция деформации по-прежнему не является непрерывной.

## 6.3. Параметрическая калибровка Платта (Platt calibration)

Изначально этот метод калибровки разрабатывался только для метода опорных векторов, оценки которого лежат на вещественной оси (по сути, это расстояния до оптимальной разделяющей классы прямой, взятые с нужным знаком). Считается, что этот метод не очень подходит для других моделей.

Предложен Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.

### 6.4. Логистическая регрессия в пространстве логитов

### 6.5. Деревья калибровки

Стандартный алгоритм строит строит суперпозицию дерева решений на исходных признаках и логистических регрессий (каждая в своем листе) над оценками алгоритма:

- Построить на исходных признаках решающее дерево (не очень глубокое),
- В каждом листе – обучить логистическую регрессию на одном признаке,
- Подрезать дерево, минимизируя ошибку.

### 6.6. Температурное шкалирование (Temperature Scaling)

Этот метод относится к классу DL-методов калибровки, так как он был разработан именно для калибровки нейронных сетей. Метод представляет собой простое многомерное обобщение шкалирования Платта.

## 7. Приемы работы с менеджером пакетов `conda`

### 7.1. Создание виртуального окружения

Создать виртуальное окружение `dashenv`

```
conda create --name dashenv
```

Создать виртуальное окружение с указанием версии `Python`

```
conda create --name testenv python=3.6
```

Создать виртуальное окружение с указанием пакета

```
conda create --name testenv scipy
```

Создать виртуальное окружение с указанием версии `Python` и нескольких пакетов

```
conda create --name testenv python=3.6 scipy=0.15.0 astroid babel
```

---

*Замечание*

Рекомендуется устанавливать сразу несколько пакетов, чтобы избежать конфликта зависимостей

---

Для того чтобы при создании нового виртуального окружения не требовалось каждый раз устанавливать базовые пакеты, которые обычно используются в работе, можно привести их список в конфигурационном файле `.condarc` в разделе `create_default_packages`

.condarc

```
ssl_verify: true
channels:
 - conda-forge
 - defaults
report_errors: true
default_python:
create_default_packages:
 - matplotlib
 - numpy
 - scipy
 - pandas
 - seaborn
```

Если для текущего виртуального окружения не требуется устанавливать пакеты из набора по умолчанию, то при создании виртуального окружения следует указать специальный флаг `--no-default-packages`

```
conda create --no-default-packages --name testenv python
```

Создать виртуальное окружение можно и из файла `environment.yml` (первая строка этого файла станет именем виртуального окружения)

environment.yml

```
name: stats2
channels:
 - conda-forge
 - defaults
dependencies:
 - python=3.6   # or 2.7
 - bokeh=0.9.2
 - numpy=1.9.*
 - nodejs=0.10.*
 - flask
 - pip:
  - Flask-Testing
```

```
conda env create -f environment.yml
```

При создании виртуального окружения можно указать путь до целевой директории, где будут размещаться файлы окружения. Следующая команда создаст виртуальное окружение в поддиректории текущей рабочей директории `envs`[4]

```
conda create --prefix ./envs jupyterlab matplotlib
```

С помощью файла спецификации можно создать *идентичное виртуальное окружение* (i) на той же платформе операционной системы, (ii) на той же машине, (iii) на какой-либо другой машине (перенести настройки окружения).

Для этого предварительно требуется создать собственно файл спецификации

```
conda list --explicit > spec-file.txt
```

Имя файла спецификации может быть любым. Файл спецификации обычно не является кросс-платформенным и поэтому имеет комментарий в верхней части файла (`#platform: osx-64`), указывающий платформу, на которой он был создан.

Теперь для того чтобы *создать* окружение достаточно воспользоваться командой

```
conda create --name myenv --file spec-file.txt
```

Файл спецификации можно использовать для установки пакетов в существующее окружение

```
conda install --name myenv --file spec-file.txt
```

## 7.2. Активация/деактивация виртуального окружения

Активировать виртуальное окружение `dashenv`

```
conda activate dashenv
```

[4] В данном случае чтобы удалить виртуальную среду достачно просто удалить директорию `envs`

Активировать виртуальное окружение в случае, когда оно создавалось с `--prefix`, можно указав полный путь до окружения

```
conda activate E:\[WorkDirectory]\[Python_projects]\directory_for_experiments\envs
```

В этом случае в строке приглашения командной оболочки по умолчанию будет отображаться полный путь до окружения. Чтобы заменить длинный префикс в имени окружения на более удобный псевдоним достаточно использовать конструкцию

```
conda config --set env_prompt ({name})
```

которая добавит в конфигурационный файл `.condarc` следующую строку

.condarc

```
...
env_prompt: ({name})
```

и теперь имя окружения будет `(envs)`.

Деактивировать виртуальное окружение

```
conda deactivate
```

## 7.3. Обновление виртуального окружения

Обновить виртуальное окружение может потребоваться в следующих случаях:

- обновилась одна из ключевых зависимостей,
- требуется добавить пакет (добавление зависимости),
- требуется добавить один пакет и удалить другой.

В любом из этих случаев все что нужно для того чтобы обновить виртуальное окружение это просто обновить файл `environment.yml`[5], а затем запустить команду

```
conda env update --prefix ./envs --file environment.yml --prune
```

Опция `--prune` приводит к тому, что `conda` удаляет все зависимости, которые больше не нужны для окружения.

## 7.4. Вывод информации о виртуальном окружении

Вывести список доступных виртуальных окружений

```
conda env list
```

Вывести список пакетов, установленных в указанном окружении

```
conda list --name myenv
```

Вывести информацию по конкретному пакету указанного окружения

```
conda list --name dashenv matplotlib
```

[5]Этот файл должен находится в той же директории что и директория окружения `envs`

### 7.5. Удаление виртуального окружения

Удалить виртуальное окружение `heroku_env`

```
conda env remove --name heroku_env
```

### 7.6. Экспорт виртуального окружения в `environment.yml`

Экспортировать активное виртуальное окружение в yml-файл

```
conda env export > environment.yml
```

## 8. Приемы работы с пакетом `Vowpal Wabbit`

## 9. Приемы работы с библиотекой `pandas`

### 9.1. Число уникальных значений категориальных признаков в объекте `DataFrame`

Для того чтобы вывести информацию по числу уникальных значений в каждом категориальном признаке некоторого объекта `pandas.DataFrame` можно воспользоваться конструкцией

```
X.select_dtypes('category').apply(lambda col: col.unique().shape[0])
```

### 9.2. Число пропущенных значений в объекте `DataFrame`

Информацию по числу пропущенных значений в каждом столбце можно вывести следующим образом

```
X.isna().any(axis=0)
```

## 10. Интерпретация моделей и оценка важности признаков с библиотекой `SHAP`

### 10.1. Общие сведения о значениях Шепли

В библиотеке `SHAP` https://github.com/slundberg/shap для оценки *важности признаков* используются *значения Шепли*[6] (Shapley value) https://en.wikipedia.org/wiki/Shapley_value.

Или несколько точнее: при построении *локальной* интерпретации (то есть интерпретации на конкретной точке данных) значения Шепли, строго говоря, оценивают *силу влияния*[7] $i$-ого признака $f_i$ на значения целевого вектора $y$, а вот *важность признака* в контексте модели можно оценить при построении *глобальной* интерпретации с помощью значений Шепли, взятых по абсолютной величине и усредненных по имеющемуся набору данных.

---

*Замечание*

Значения Шепли объясняют как «справедливо» оценить вклад каждого признака в прогноз модели

---

[6] Термин пришел из теории кооперативных игр

[7] Еще эту оценку можно интерпретировать как *вклад*

Значения Шепли $i$-ого признака на *конкретном объекте* (на текущей точке данных) вычисляются следующим образом (здесь сумма распространяется на все подмножества признаков $S$ из множества признаков $N$, не содержащие $i$-ого признака)

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} \Big( \underbrace{v(S \cup \{i\}) - v(S)}_{f_i - contribution} \Big),$$

где $n$ – общее число признаков; $v(S \cup \{i\})$ – прогноз модели с учетом $i$-ого признака; $v(S)$ – прогноз модели без $i$-ого признака.

Выражение $v(S \cup \{i\}) - v(S)$ – это вклад $i$-ого признака. Если теперь вычислить среднее вкладов по всем возможным перестановкам, то получится «честная» оценка вклада $i$-ого признака.

Значение Шепли для $i$-ого признака вычисляется для каждой точки данных (например, для каждого клиента в выборке) на всех возможных комбинациях признаков (в том числе и для пустых подмножеств $S$).

---

*Замечание*

Метод анализа важности признаков, реализованный в библиотеке `SHAP`, является и *согласованным*, и *точным* (см. Interpretable Machine Learning with XGBoost)

---

## 10.2. Пример построения локальной и глобальной интерпретаций

Примеры использования библиотеки `SHAP` не только для tree-base моделей можно найти по адресу https://github.com/slundberg/shap/tree/master/notebooks/tree_explainer.

Решается задача регрессии для классического набора данных `boston`. Требуется предсказать стоимость квартиры.

```
import shap
import os
import pandas as pd
import numpy as np
from pandas import DataFrame, Series
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_boston
#%matplotlib inline  # если код оформляется в JupyterLab
#shap.initjs()  # если код оформляется в JupyterLab

boston = load_boston()
X, y = boston['data'], boston['target']  # numpy-массивы

# объекты pandas
X_full = DataFrame(X, columns=boston['feature_names'])
y_full = Series(y, name = 'PRICE')

X_train, X_test, y_train, y_test = train_test_split(X_full, y_full, random_state=42)

rf = RandomForestRegressor(n_estimators=500).fit(X_train, y_train)

explainer = shap.TreeExplainer(rf)  # <- NB
shap_values_train = explainer.shap_values(X_train)  # <- NB
```

### 10.2.1. Локальная интерпретация отдельной точки данных обучающего набора

Теперь можно построить локальную интерпретацию для одной точки данных из обучающего набора (см. рис. 6)

К вопросу о локальной интерпретации отдельной точки данных обучающего набора

```
row = 1
shap.force_plot(
    explainer.expected_value,  # ожидаемое значение
    shap_values_train[row, :],  # 2-ая строка в матрице значений Шепли
    X_train.iloc[row, :]  # 2-ая строка в обучающем наборе данных
)
```

Можно считать, что `explainer.expected_value` это значение, полученное усреднением целевого вектора по точкам обучающего набора данных, т.е. `y_train.mean()`.


Рис. 6. Локальная интерпретация для одной точки данных обучающего набора

Еще можно построить график частичной зависимости (рис. 7)

```
shap.dependence_plot('LSTAT', shap_values, X_train)
```

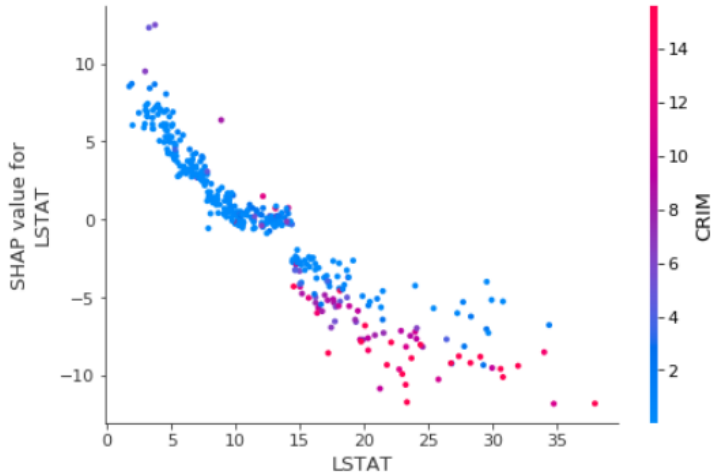


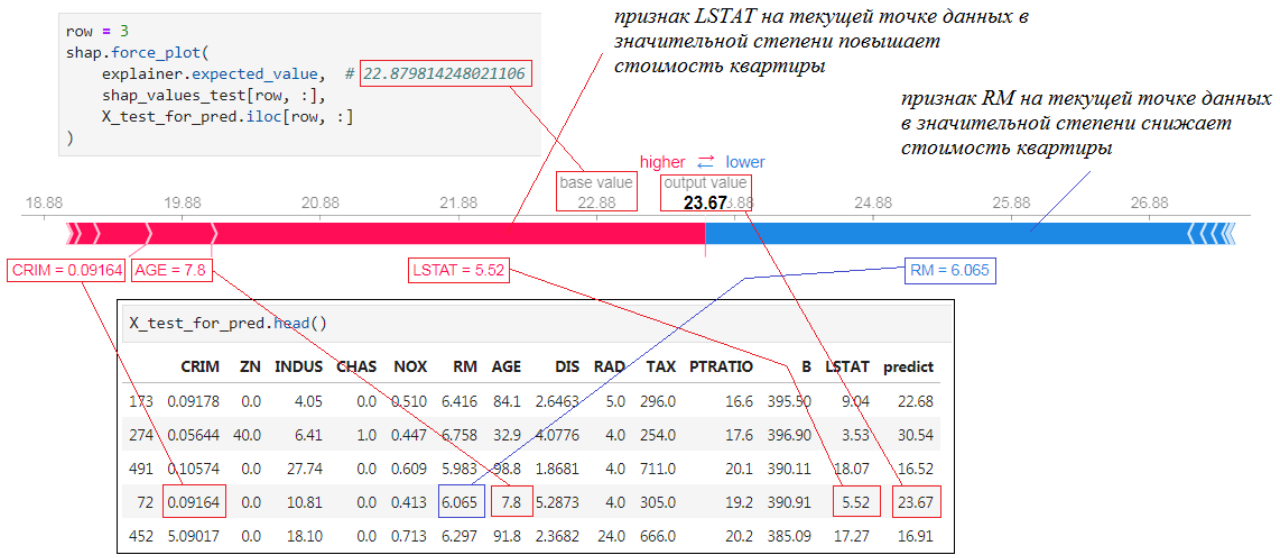

Рис. 7. График частичной зависимости признака `LSTAT` от значений Шепли с учетом влияния признака `CRIM`

### 10.2.2. Локальная интерпретация отдельной точки данных тестового набора

Прежде чем приступить к вычислению значений Шепли, следует создать поверхностную копию тестового набора данных

```
X_test_for_pred = X_test.copy()
X_test_for_pred['predict'] = np.round(rf.predict(X_test), 2)

explainer = shap.TreeExplainer(rf)
# вычисляем значения Шепли для тестового набора данных со столбцом 'predict'
```




| | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | predict |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 173 | 0.09178 | 0.0 | 4.05 | 0.0 | 0.510 | 6.416 | 84.1 | 2.6463 | 5.0 | 296.0 | 16.6 | 395.50 | 9.04 | 22.68 |
| 274 | 0.05644 | 40.0 | 6.41 | 1.0 | 0.447 | 6.758 | 32.9 | 4.0776 | 4.0 | 254.0 | 17.6 | 396.90 | 3.53 | 30.54 |
| 491 | 0.10574 | 0.0 | 27.74 | 0.0 | 0.609 | 5.983 | 98.8 | 1.8681 | 4.0 | 711.0 | 20.1 | 390.11 | 18.07 | 16.52 |
| 72 | 0.09164 | 0.0 | 10.81 | 0.0 | 0.413 | 6.065 | 7.8 | 5.2873 | 4.0 | 305.0 | 19.2 | 390.91 | 5.52 | 23.67 |
| 452 | 5.09017 | 0.0 | 18.10 | 0.0 | 0.713 | 6.297 | 91.8 | 2.3682 | 24.0 | 666.0 | 20.2 | 385.09 | 17.27 | 16.91 |

Рис. 8. Локальная интерпретация для одной точки данных тестового набора

```
shap_values_test = explainer.shap_values(X_test_for_pred)
```

Теперь можно построить локальную интерпретацию для отдельной точки данных тестового набора (рис. 8).

Из рис. 8 видно, что признаки с различной «силой»[8], которая определяется значениями Шепли, смещают предсказание модели на данной точке. Например, признак `LSTAT` (процент населения с низким социальным статусом) в значительной степени *повышает*[9] стоимость квартиры на данной точке по отношению к базовому значению `base_value`, а признак `RM` (среднее число комнат в жилом помещении) в значительной степени снижает.

К вопросу о локальной интерпретации отдельной точки данных тестового набора

```
row = 3
shap.force_plot(
    explainer.expected_value,  # 22.879814248021106
    #y_train.mean()  # 22.907915567282323
    shap_values_test[row, :],
    X_test_for_pred.iloc[row, :]
)
```

### 10.2.3. Глобальная интерпретация модели на тестовом наборе данных

Удобно работать с диаграммой рассеяния `shap.summary_plot` (рис. 9), на которой изображаются признаки в порядке убывания их важности, с одновременным указанием того, насколько сильно каждый из признаков влияет на целевую переменную.

```
shap.summary_plot(shap_values_test, X_test_for_pred)
```

Какие выводы можно сделать из рис. 9:

- Признаки `LSTAT`, `RM` и `CRIM` имеют высокую важность для модели в целом,

[8] Ширина полосы

[9] Потому что значение этого признака невелико; чем меньше процент населения с низким социальным статусом проживает в округе, тем выше стоимость квартиры

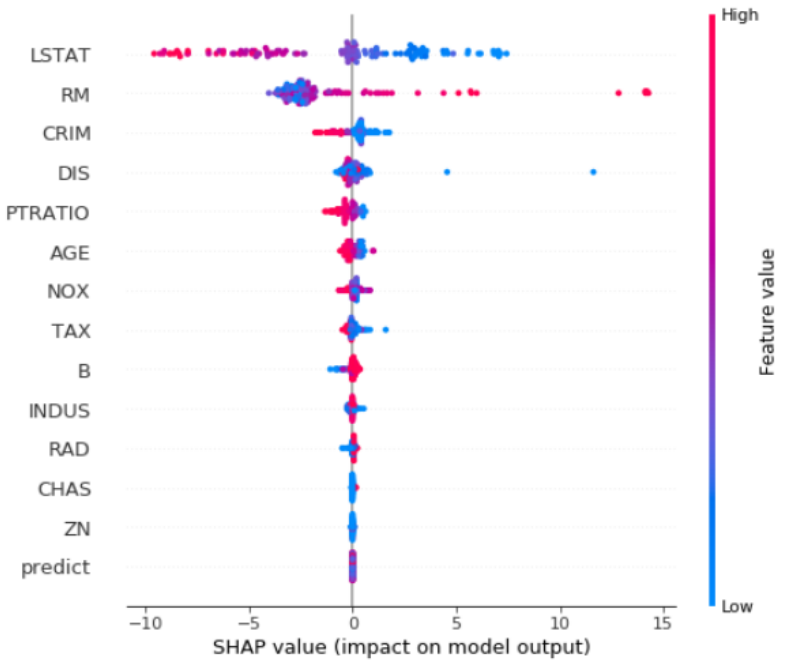



Рис. 9. Диаграмма рассеяния для точек тестового набора данных

- Для признака `LSTAT` наблюдается отрицательная статистическая зависимость от целевой переменной, т.е. низкие значения этого признака отвечают высоким значениям целевой переменной (стоимости на квартиру),
- Для признака `RM` наблюдается положительная статистическая зависимость от целевой переменной: чем больше комнат в жилом помещении, тем выше стоимость квартиры.

Затем можно детальнее изучить графики частичной зависимости, построенные на тестовом наборе данных. Рассмотрим зависимость признака `CRIM` (уровень преступности в городе на душу населения) от значений Шепли, вычисленных для этого признака (рис. 10).

```
shap.dependence_plot('CRIM', shap_values_test[:, :-1], X_test_pred.iloc[:, :-1])
```

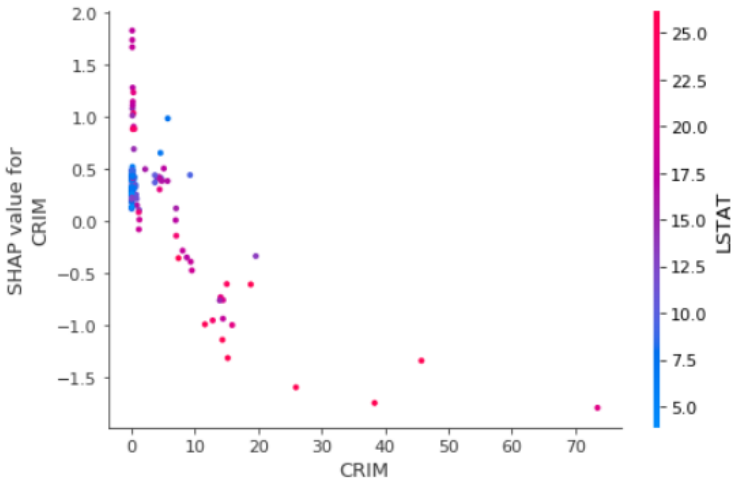



Рис. 10. График частичной зависимости признака `CRIM` от значений Шепли с учетом влияния `LSTAT`

Какие выводы можно сделать из рис. 10:

- Чем выше уровень преступности в городе, тем в большей степени снижается стоимость квартиры,
- Не везде, где проживает высокий процент населения с низким социальным статусом наблюдается высокий уровень преступности, однако в тех местах, где регистрируется высокий уровень преступности одновременно регистрируется и высокий процент населения с низким социальным статусом.

## 11. Перестановочная важность признаков в библиотеке `eli5`

Еще важность признаков можно оценивать с помощью так называемой *перестановочной важности* (permutation importances) https://www.kaggle.com/dansbecker/permutation-importance.

Идея проста: нужно в заранее отведенном для исследования важности признаков наборе данных (валидационном наборе) перетасовать значения признака, влияние которого изучается на данной итерации, оставив остальные признаки (столбцы) и целевой вектор без изменения.

Признак считается «важным», если метрики качества модели падают, и соответственно – «неважным», если перестановка не влияет на значения метрик. Перестановочная важность вычисляется после того как модель будет обучена.

---

*Замечание*

Перестановочная важность обладает свойством *согласованности*, но не обладает свойством *точности*

Interpretable Machine Learning with XGBoost

---

Рассмотрим задачу построения регрессионной модели на наборе данных `load_boston`

```
import eli5
import pandas as pd
from eli5.sklearn import PermutationImportance
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_boston
from pandas import DataFrame, Series

boston = load_boston()

X_train, X_test, y_train, y_test = train_test_split(boston['data'],
                                                    boston['target'],
                                                    random_state=2)

X_train_sub, X_valid, y_train_sub, y_valid = train_test_split(X_train,
                                                              y_train,
                                                              random_state=0)

# модель случайного леса, как обычно, обучается на обучающей выборке
rf = RandomForestRegressor(n_estimators=500).fit(X_train_sub, y_train_sub)

# модель перестановочной важности обучается на валидационном наборе данных
perm = PermutationImportance(rf, random_state=42).fit(X_valid, y_valid)

eli5.show_weights(perm, feature_names = boston['feature_names'])  # визуализирует перестановочны
    е важности признаков
```

## Список литературы


1. *Лутц М.* Изучаем Python, 4-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 1280 с.
2. *Бизли Д.* Python. Подробный справочник. – Пер. с англ. – СПб.: Символ-Плюс, 2010. – 864 с.
3. *Чакон С.*, *Штрауб Б.* Git для профессионального программиста. – СПб.: Питер, 2020. – 496 с.
4. *Рамальо Л.* Python. К вершинам мастерства. – М.: ДМК Пресс, 2016. – 768 с.