Лабораторная работа №2: Линейные модели. Кросс-валидация.

Упражнение 2

- 1. Данные своего варианта (см. таблицу ниже) разделить на выборку для построения моделей (80%) и отложенные наблюдения (20%). Оставить в таблице только указанные в варианте переменные. Отложенные наблюдения использовать только в задании 6.
- 2. Провести предварительный анализ данных с помощью описательных статистик и графиков, оценить взаимосвязь.
- 3. Проверить Y на нормальность. Если он распределён не по нормальному закону, прологарифмировать и снова провести анализ взаимосвязей переменных.
- 4. Составить список возможных спецификаций моделей множественной регрессии (на исходной Y и на логарифме Y).
- 5. Оценить параметры моделей из списка. Оценить точность моделей методом перекрёстной проверки, указанным в варианте. Найти самую точную из моделей для Y . Найти самую точную из моделей для log(Y).
- 6. Сделать прогноз с помощью самых точных моделей на отложенные наблюдения. Рассчитать MSE_{test} вручную и выбрать одну наиболее точную модель. Проинтерпретировать её параметры.

Варианты

Номер варианта – номер студента в списке. Студент под номером 21 берёт вариант 1, под номером 22 – 2, и т.д.

В качестве ядра генератора случайных чисел (в частности, для разделения данных на выборку для построения моделей и отложенные наблюдения) используйте номер своего варианта.

Наборы данных и справочники к ним выложены по адресу: https://qithub.com/ania607/ML/tree/main/data .

Вариант	Проверка	Набор данных	Зависимая	Объясняющие переменные	
			переменная	непрерывные	фиктивные
1	LOOCV	Boston_for_lab	medv /	zn / доля земли	tax_over_400
			медианная	под жилую	/ 1 если полная
			стоимость	застройку	ставка налога на
			домов, тыс.	indus / доля	имущество на
			долл. США	акров, не	\$10000
				связанных с	превышает 400

	1	T	1	1	
				розничной	
				торговлей, на	
				город	
2	K-	Boston for lab	medv /	rm / среднее	tax over 400
2	VAL(10)			_	/ 1 если полная
	VAL(10)		медианная	количество комнат	·
			стоимость	в доме пох /	ставка налога на
			домов, тыс. долл. США	концентрация	имущество на \$10000
			долл. Спи	оксидов азота	превышает 400
				(частей на 10	превышает 400
				миллионов)	
3	K-	Boston for lab	medv /	rm / среднее	tax over 400
	VAL(5)		медианная	количество комнат	/ 1 если полная
			стоимость	в доме	ставка налога на
			домов, тыс.	dis /	имущество на
			долл. США	средневзвешенное	\$10000
				расстояние до	превышает 400
				пяти бостонских	_
				центров занятости	
4	LOOCV	Boston_for_lab	medv /	rm / среднее	tax_over_400
			медианная	количество комнат	/ 1 если полная
			стоимость	в доме	ставка налога на
			домов, тыс.	indus / доля	имущество на
			долл. США	акров, не	\$10000
				связанных с	превышает 400
				розничной	
				торговлей, на	
5	K-	Boston for lab	medv /	город indus / доля	tax over 400
3	VAL(10)	Doscon_ror_rap	медианная	акров, не	/ 1 если полная
	(()		стоимость	связанных с	ставка налога на
			домов, тыс.	розничной	имущество на
			долл. США	торговлей, на	\$10000
				город	превышает 400
				crim / уровень	
				преступности на	
				душу населения	
				по району	
6	K-	Auto_for_lab	mpg /	displacement /	cyl_over_4/
	VAL(5)		пробег	объемн двигателя	1 если число
			автомобиля	(в кубических	цилиндров
			на галлоне	дюймах)	больше 4
			топлива	acceleration /	
				время разгона с 0	
				до 60 миль в час (в	
7	LOOCV	Auto for lab	mpg /	секундах) horsepower /	cyl over 4/
/	1000	11400_101_140	пробег	мощность (в	1 если число
			автомобиля	лошадиных силах)	цилиндров
			на галлоне	weight / macca	больше 4
			топлива	(в фунтах)	
8	K-	Auto for lab	mpg /	displacement /	cyl over 4/
	VAL(10)		пробег	объем двигателя (в	1 если число
			автомобиля	кубических	цилиндров
			на галлоне	дюймах)	больше 4
			топлива	<u> </u>	

				weight / масса	
				(в фунтах)	
9	K- VAL(5)	Auto_for_lab	тробег автомобиля на галлоне топлива	acceleration / время разгона с 0 до 60 миль в час (в секундах) horsepower / мощность (в лошадиных силах)	cyl_over_4/ 1 если число цилиндров больше 4
10	LOOCV	Carseats	Sales / продажа (в тысячах штук) в каждом магазине	Price / цены компании на автокресла в каждом магазине Advertising / бюджет затрат на рекламу в каждом магазине (в тысячах долларов)	ShelveLoc/ качество стеллажа для размещения автокресел в каждом магазине: Плохое, Хорошее и Среднее
11	K- VAL(10)	Carseats	Sales / продажа (в тысячах штук) в каждом магазине	Price / цены компании на автокресла в каждом магазине Population / плотность населения	ShelveLoc/ качество стеллажа для размещения автокресел в каждом магазине: Плохое, Хорошее и Среднее
12	K- VAL(5)	Carseats	Sales / продажа (в тысячах штук) в каждом магазине	Price / цены компании на автокресла в каждом магазине Income / уровень дохода сообщества (в тысячах долларов)	ShelveLoc/ качество стеллажа для размещения автокресел в каждом магазине: Плохое, Хорошее и Среднее
13	LOOCV	Carseats	Sales / продажа (в тысячах штук) в каждом магазине	Price / цены компании на автокресла в каждом магазине CompPrice / цена конкурента в каждом магазине	ShelveLoc/ качество стеллажа для размещения автокресел в каждом магазине: Плохое, Хорошее и Среднее
14	K- VAL(10)	College_for_lab	Grad_Rate / выпускной балл	Тор10регс / процент зачисленных студентов, которые в старшей школе относились к топ-10% в диапазоне по освоению	Private / частный или государственный университет: Да - частный, Нет - государственный

	1	T	I	I — 1	1
				F_Undergrad /	
				количество	
				студенческой	
				формы обучения	
15	K-	College_for_lab	Grad_Rate	F_Top25perc /	Private /
	VAL(5)		/	процент	частный или
			выпускной	зачисленных	государственный
			балл	студентов,	университет: Да
				которые в старшей	- частный, Нет -
				школе относились	государственный
				к топ-25% в	тобударотвонным
				диапазоне по	
				освоению	
				F Undergrad /	
				_	
				количество	
				студенческой	
				формы обучения	
16	LOOCV	College_for_lab	Grad_Rate	Accept /	Private /
			/	количество	частный или
			выпускной	реализованных	государственный
			балл	заявок на	университет: Да
				поступление	- частный, Нет -
				Expend /	государственный
				расходы на	
				обучение на	
				одного студента	
17	K-	College for lab	Grad Rate	Accept /	Private /
	VAL(10)		_	количество	частный или
			выпускной	реализованных	государственный
			балл	заявок на	университет: Да
				поступление	- частный, Нет -
				Top10perc /	государственный
				процент	тосударственным
				зачисленных	
				студентов,	
				которые в старшей	
				школе относились	
				к топ-10% в	
				диапазоне по	
10	77		G 1 D 1	освоению	D ' /
18	K-	College_for_lab	Grad_Rate	Expend /	Private /
	VAL(5)		/	расходы на	частный или
			выпускной	обучение на	государственный
			балл	одного студента	университет: Да
				Top25perc /	- частный, Нет -
				процент	государственный
				зачисленных	
				студентов,	
				которые в старшей	
				школе относились	
				к топ-25% в	
				диапазоне по	
				освоению	
19	LOOCV	College for lab	Grad Rate	Expend /	Private /
17		1	/	расходы на	частный или
			выпускной	обучение на	государственный
			балл	•	
			Oalli	одного студента	университет: Да
]			P_Undergrad /	- частный, Нет -

				количество студентов, обучающихся по совместительству	государственный
20	K- VAL(10)	College_for_lab	Grad_Rate / выпускной балл	Ассерт / количество реализованных заявок на поступление F_Undergrad / количество студенческой формы обучения	Private / частный или государственный университет: Да - частный, Нет - государственный